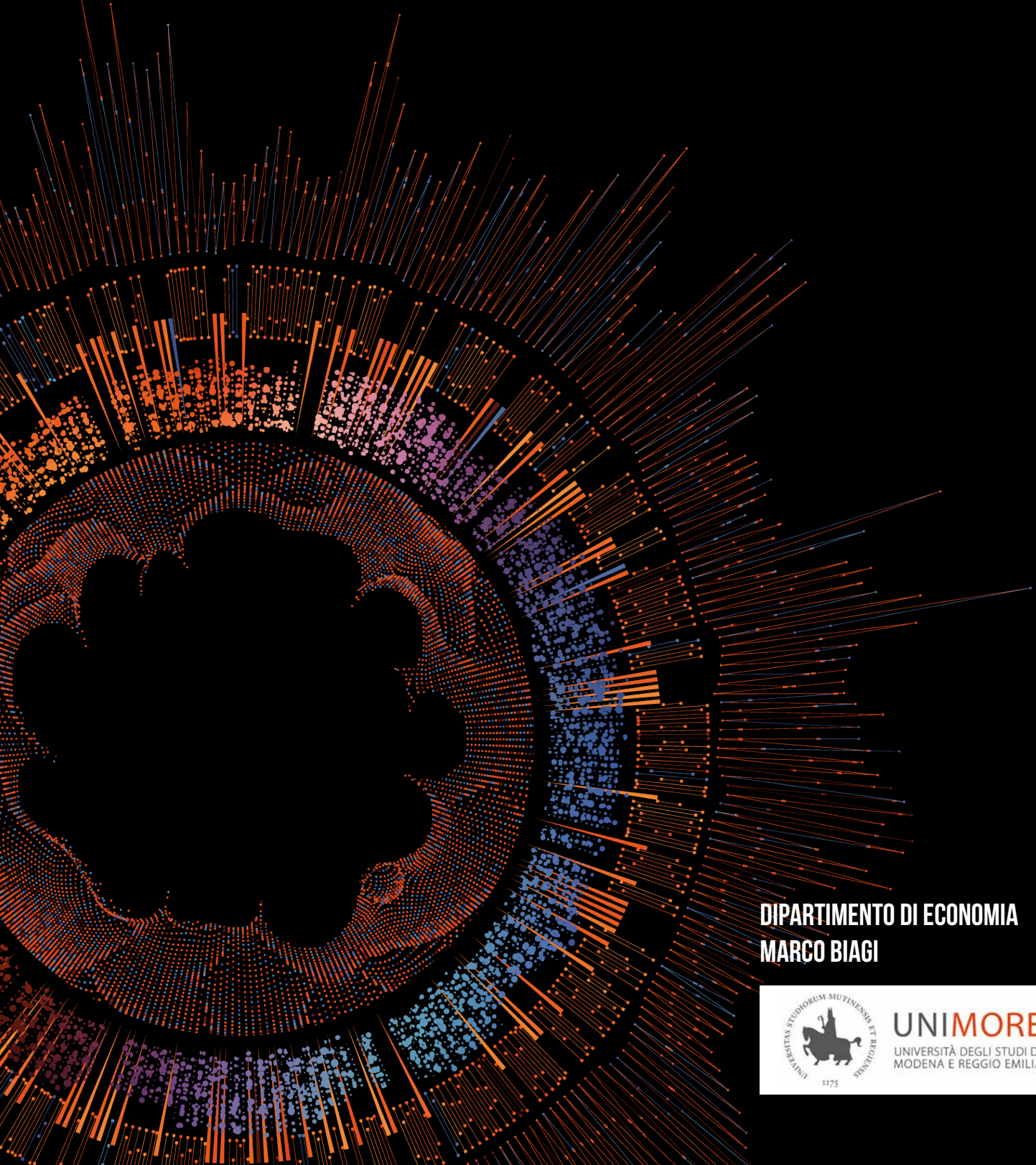
CAVICCHIOLI M. • DEMARIA F. • FRANCO VILLORIA M. • FREDERIC P. • MORLINI I.

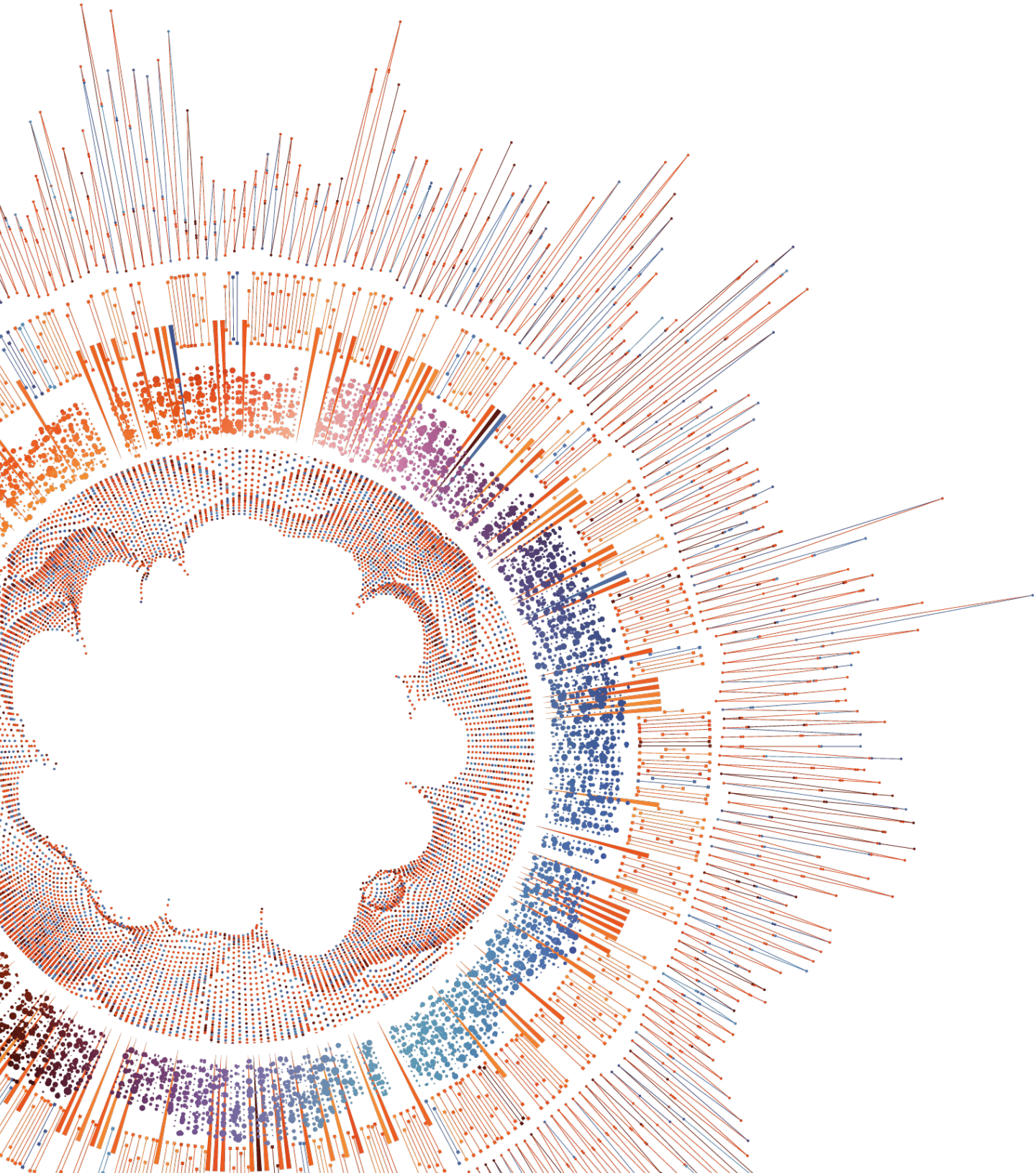
STATISTICA:
LA SCIENZA CHE MODELLA I DATI

UN'INTRODUZIONE ALLE DIVERSE TIPOLOGIE DI DATI



DIPARTIMENTO DI ECONOMIA
MARCO BIAGI

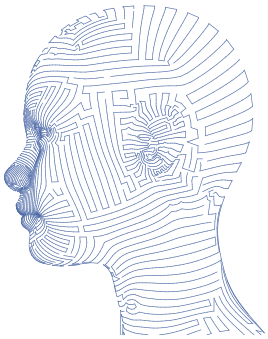
UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA



SOMMARIO

PREFAZIONE

di Monica Prandini



Dati, dati e ancora *Dati, Open data, Big Data:* parole *magiche* del nostro secolo che giocano un ruolo da protagonista. Oggi è estremamente più semplice raccogliere informazioni: si pensi solo che negli anni Ottanta i dati relativi ai censimenti venivano raccolti a mano e poi inseriti da persone negli appositi software per essere rielaborati. Con Internet alla portata di tutti, i dati possono, ad esempio, essere raccolti con un semplice click quando ci si collega ad un sito.

Banche, trasporti, ambiente, medicina e sanità sono solo alcuni dei settori in cui i dati giocano un ruolo di importanza fondamentale. Ne faccio solo alcuni esempi.

Le Pubbliche Amministrazioni stanno investendo in progetti che rendano accessibili gli Open Data. Basti pensare al sito che sta sviluppando il Comune di Bologna grazie al progetto *"Per una governance collaborativa dei dati della comunità"*, realizzato in partnership con la città di Barcellona nell'ambito di un protocollo di collaborazione tra le due municipalità siglato già nel 2018.

Altro tema attuale è rappresentato dal riscaldamento globale che troviamo come *obiettivo 13* nell'agenda 2023. È proprio grazie all'analisi dei dati dei quali siamo in possesso che si possono ricercare strategie per affrontare la situazione. Come si legge sul sito dell'*Intergovernmental Panel on Climate Change "the IPCC was created to provide policymakers with regular scientific assessments on climate change, its implications and potential future risks, as well as to put forward adaptation and mitigation options"*.

Nella sezione *Science & Math* del *The New York Times - The learning network: lesson plans and Teaching Ideas* troviamo, ad esempio, varie proposte relative all'analisi di dati e di grafici relativi ad argomenti di attualità.

La Generazione Z - i cosiddetti *nativi digitali* - ragazzi che, sin dall'infanzia hanno potuto vivere usando tecnologia e social media, sono però in grado di capire, gestire e interpretare queste enormi quantità di dati?

Sono una docente di scuola secondaria superiore, che insegna matematica/fisica/informatica da una trentina di anni, convinta che l'analisi dei dati rappresenti una delle chiavi per leggere e interpretare il futuro.

Nella scuola attuale, di solito, la statistica non viene ancora vissuta dai docenti come un argomento di fondamentale importanza nella formazione di uno studente al pari dell'algebra o della geometria, ma viene spesso lasciata da trattare "qualora rimanga tempo". Ritengo invece che sia fondamentale dare uno spazio maggiore a questa scienza per permettere ai nostri giovani di essere preparati alle richieste del mercato. Per questo motivo ho iniziato con una mia classe del triennio a parlare di Analisi dei Dati all'inizio del 2018. Con l'arrivo della pandemia nel 2020 il dato ha regnato sovrano e il mio obiettivo è sempre stato quello di far capire agli studenti che il dato grezzo non fornisce informazioni. Anche la semplice lista dei voti presi in una materia, elencati singolarmente, non racconta una storia.

Partendo da questo e dalla sempre maggior importanza assegnata al dato in tantissimi settori, ho iniziato a proporre nelle mie classi un percorso che si basi sull'analisi dei dati. Il mio approccio all'analisi dei dati è sempre stato incentrato sul *learning by doing*. In particolare, lo scorso anno, nella mia scuola, in alcune classi del biennio, è stato proposto un progetto in collaborazione con il Dipartimento di Economia Marco Biagi di UNIMORE articolato in diversi incontri e basato sull'approccio statistico e probabilistico del dato. L'interesse e il coinvolgimento

da parte delle classi è stato notevole: l'incontro per i ragazzi è stato molto interessante e il loro feedback positivo ha spinto noi docenti di matematica a richiedere nuovamente al Dirigente la possibilità di riproporre, qualora possibile, l'intervento del team di docenti di Statistica DEMB di Unimore nelle classi del biennio.

Posso affermare che i ragazzi sono realmente interessati a capire come analizzare e interpretare i dati. Per loro il percorso non è semplice, perché non sono abituati a studiare in questo modo: una buona modellizzazione e astrazione richiede competenze di livello superiore che gli studenti non posseggono nemmeno all'inizio del triennio e quindi vanno guidati ed educati alla cultura del dato. Da sottolineare è sicuramente il fatto che i ragazzi in classe seguono con la massima attenzione e partecipazione anche senza dover far fare loro le verifiche canoniche per "obbligarli" a studiare in quanto capiscono quanto importante sia per il loro futuro possedere competenze in questo campo: la modellizzazione e la loro applicazione in diversi ambiti sociali offrono loro non solo competenze tecniche spendibili per il loro futuro, ma anche l'acquisizione di quelle *soft skills*, di quelle abilità sociali e interpersonali come la capacità di fare gioco di squadra lavorando in team, il potenziamento delle abilità comunicative e l'ottimizzazione nella gestione del tempo spendibili in ogni settore.

L'attuale Ministro dell'Istruzione e del Merito Giuseppe Valditara afferma, parlando agli studenti dell'istituto 'Opere sociali Don Bosco Salesiani' a Sesto San Giovanni, che "la vera sfida che noi vogliamo lanciare come Ministero e come Governo per riformare l'insegnamento delle Stem che ci vede, purtroppo, oggi molto indietro rispetto ad altri Paesi europei è proprio quella di partire dalla realtà per arrivare alle astrazioni. [...] quindi bisognerà cambiare un po' anche l'insegnamento della matematica".

Anche il Presidente della Repubblica Sergio Mattarella nel discorso di fine 2022 afferma che "l'altro cambiamento che stiamo vivendo, e di cui probabilmente fatichiamo tuttora a comprendere la portata, riguarda la trasformazione digitale. L'uso delle tecnologie digitali ha già modificato le nostre vite, le nostre abitudini e probabilmente i modi di pensare e vivere le relazioni interpersonali. Le nuove generazioni vivono già pienamente questa nuova dimensione. La quantità e la qualità dei dati, la loro velocità possono essere elementi posti al servizio della crescita delle persone e delle comunità. Possono consentire di superare arretratezze e divari, semplificare la vita dei cittadini e modernizzare la nostra società. Occorre compiere scelte adeguate, promuovendo una cultura digitale che garantisca le libertà dei cittadini".

Da qui l'importanza fondamentale di poter essere supportati durante il percorso scolastico da testi come questo che trattano argomenti imprescindibili per il nostro tempo insieme alla realizzazione di progetti di Public Engagement dell'Università nei confronti dei giovani per avvicinarli maggiormente sia al mondo universitario che a quello del lavoro. Inoltre, il contatto diretto con docenti universitari gioca un ruolo importante per aumentare l'interesse dei ragazzi perché permette loro di confrontarsi in modo diretto con un mondo che per loro rappresenta il futuro.

Concludo riportando le parole del Presidente Sergio Mattarella: "Il terzo grande investimento sul futuro è quello sulla scuola, l'università, la ricerca scientifica. È lì che prepariamo i protagonisti del mondo di domani. Lì che formiamo le ragazze e i ragazzi che dovranno misurarsi con la complessità di quei fenomeni globali che richiederanno competenze adeguate, che oggi non sempre riusciamo a garantire. Il Piano Nazionale di Ripresa e Resilienza spinge l'Italia verso questi traguardi. Non possiamo permetterci di perdere questa occasione. Lo dobbiamo ai nostri giovani e al loro futuro".

INTRODUZIONE

Se c'è una cosa che l'emergenza sanitaria in seguito alla diffusione del coronavirus ha messo in evidenza in tutto il mondo è la centralità e l'importanza dei dati per monitorare gli eventi, capire le dinamiche e prendere decisioni.

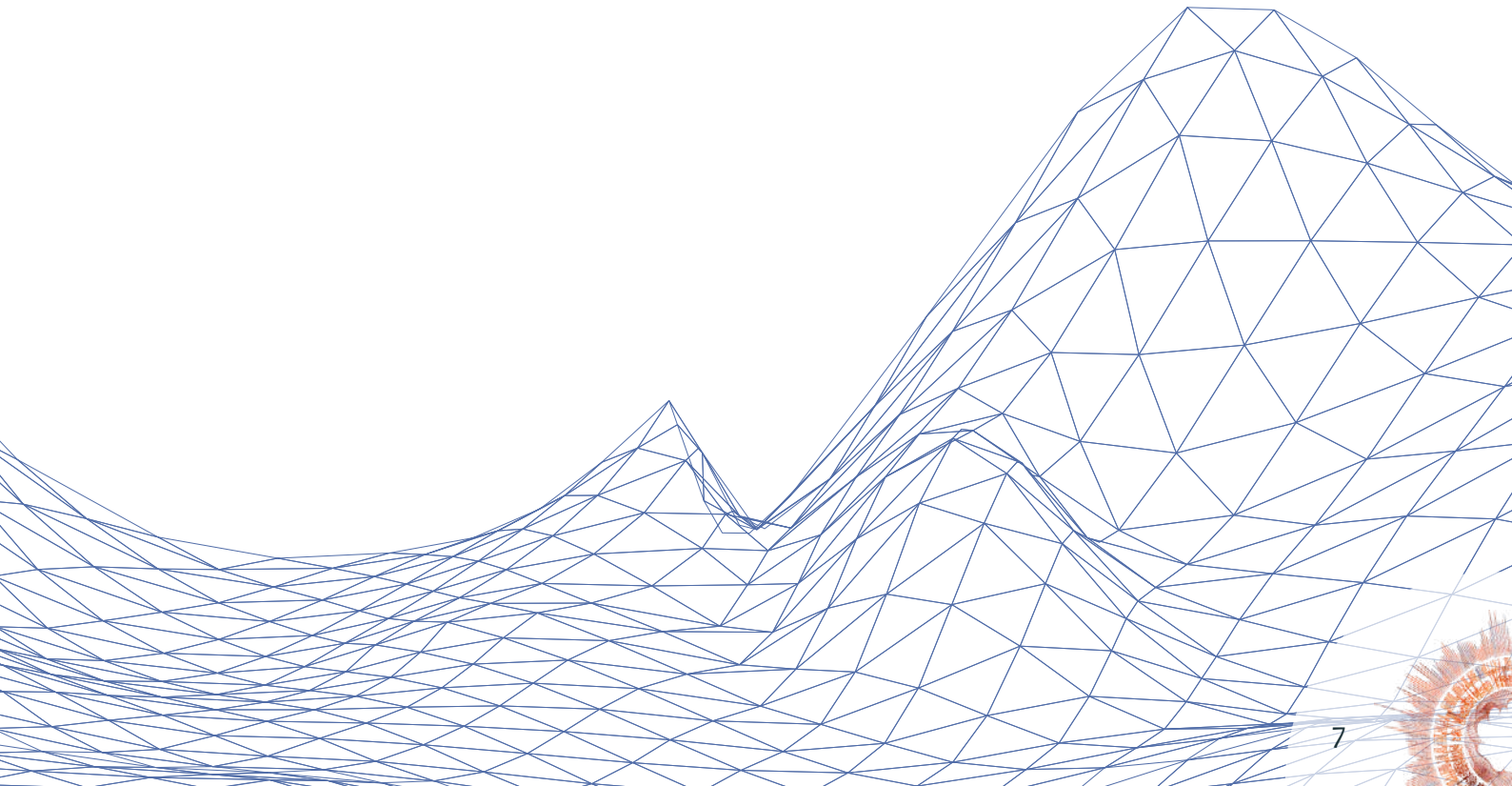
Al tempo stesso, la pandemia ha accelerato la trasformazione digitale e il processo di datificazione in molti ambiti della vita individuale e sociale di diversi soggetti economici. In particolare, il processo di datificazione ha permesso di trasformare aspetti della nostra esistenza e delle nostre azioni in dati. Tale processo si è articolare nel raccogliere dati, digitalizzarli ed analizzarli per trasformarli in informazioni utili e, a volte, dotate di valore economico.

La scienza che è fondamento all'analisi dei dati è la statistica. Le metodologie statistiche che trasformano i dati in informazioni vengono applicate a vari ambiti della vita sociale: salute, risorse alimentari, crescita e distribuzione della ricchezza, violenza, diritti, guerre, cultura, consumo di energia, istruzione, e cambiamento ambientale, e altri.

Per far comprendere altri aspetti di questa disciplina, questa pubblicazione raccoglie i contributi di docenti universitari di statistica di Unimore i quali raccontano i processi e le metodologie con un linguaggio semplice e diretto.

Il libro nasce da un'iniziativa di Public Engagement promossa dall'Università di Modena e Reggio Emilia (UNIMORE) ed è legata ad una precedente attività che ha visto la pubblicazione di un primo volume dal titolo "*Che cos'è la statistica? Una prima introduzione alla scienza dei dati*" [ISBN: 978-88-89427-03-3].

Questo secondo volume si propone di illustrare l'importanza della scienza statistica nel modellare diverse tipologie di dati, dai dati storici ai dati sociali, dai dati spaziali ed ambientali ai cosidetti "big data" e di condividere i fondamentali dei modelli anche con gli studenti più giovani, particolarmente negli anni precedenti al percorso universitario.



CAPITOLO 1. LE DISTRIBUZIONI STATISTICHE

Scrivono Agresti, e Franklin, (2007), nel loro celebre libro:

> *"Statistics is the art and science of designing studies and analyzing the data that those studies produce. Its ultimate goal is translating data into knowledge and understanding of the world around us. In short,* ***statistics is the art and science of learning from data****"*[1].
>
> *La statistica è l'arte e la scienza di pianificare la raccolta e l'analisi dei dati che tali ricerche producono. Il suo fine è di trasformare i dati in conoscenza e comprensione del mondo circostante. In sintesi,* ***la statistica è l'arte e la scienza di imparare dai dati****.*

La statistica è dunque la scienza che si occupa dei dati: la loro definizione, la loro raccolta, l'analisi e le conclusioni di tali analisi.

1.1 I DATI, UNITÀ E POPOLAZIONE STATISTICA

Un dato, dal latino *Datum*, è un fatto osservato (è dato). Per esempio:

- Oggi Gianni ha preso un voto sufficiente in italiano.
- Michele, ieri, è stato ricoverato all'Ospedale di Baggiovara.
- La settimana scorsa la BCE ha alzato i tassi di interesse di mezzo punto.
- Il Prodotto Interno Lordo (PIL) dell'Italia nel 2022.

I dati sono collezioni di fatti, per esempio:

- Il voto in matematica del primo quadrimestre di tutti gli studenti del plesso.
- L'età di ogni insegnante delle scuole superiori della provincia di Modena.
- La differenza tra nuovi ricoveri e dimissioni dell'Ospedale di Baggiovara per tutti i giorni del 2022.
- Il PIL italiano dal 2012 al 2022.

I dati vengono raccolti su individui (**unità statistiche**), che sono accomunate per alcuni aspetti e presentano *variabilità* su altri, creando così un *fenomeno collettivo*.

La Statistica è dunque la scienza che studia i dati che definiscono i fenomeni collettivi e che presentano forme di variabilità. La Statistica individua: concetti, metodi, e strumenti per la loro analisi. L'unità statistica è l'elemento su cui si osservano le caratteristiche oggetto di studio. Una **popolazione statistica** o collettivo statistico è un insieme di *unità statistiche omogenee* rispetto a una o più caratteristiche o caratteri.

Esempio >>> Variabile = genere (Maschio, Femmina)

- Unità statistica = il singolo individuo in quest'aula
- Popolazione = gli studenti di quest'aula.

Esempio >>> Variabile = stato (difettoso, non difettoso)

- Unità = pezzo prodotto
- Popolazione = tutti i pezzi prodotti da settembre 2019

Esempio >>> Variabile = numero giorni di degenza (1, 2, 3, 4, …)

1 Agresti, A., and Franklin, C. (2007), Statistics: The Art and Science of Learning from Data, Upper Saddle River, Pearson Prentice Hall.

- Unità = individuo ricoverato
- Popolazione = tutti i ricoverati dell'ospedale XXX dal 2012 al 2020

1.2 CODIFICARE I DATI

I fatti (dati) che possiamo osservare nel mondo intorno a noi sono tanti e molto diversi gli uni dagli altri, mentre i fatti nel linguaggio naturale si esprimono con frasi, i dati in statistica, per necessità di sintesi, vanno codificati. Se stiamo collezionando il sesso di un gruppo di persone, anziché scrivere

la prima persona intervistata è di sesso maschile,

scriveremo

$x_1 = M_1$,

dopo avere stabilito convenzionalmente di attribuire *M* ai maschi e *F* alle femmine. La *x* è una lettera di comodo che si usa per indicare la caratteristica che stiamo esaminando (in questo caso il sesso), al piede il numero indica l'ordine in cui è stato rilevato. Immaginando di avere intervistato 5 individui otterremmo, per esempio

$$x_1 = M, x_2 = F, x_3 = F, x_4 = M, x_5 = M$$

Che tradotto in linguaggio naturale significa: il primo individuo intervistato è maschio, la seconda femmina, ecc. Sapere il numero di materie insufficienti per ogni studente ha modalità di codifica diversa e usiamo i numeri, anziché scrivere

la prima persona intervistata non ha alcuna materia insufficiente

scriveremo

$x_1 = 0$,

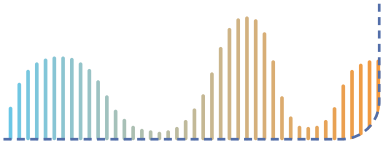
dopo avere stabilito convenzionalmente cha la *x* è il numero di materie insufficienti e che si esprime con un numero intero. Immaginando di avere intervistato cinque individui otterremmo, per esempio

$$x_1 = 0, x_2 = 2, x_3 = 0, x_4 = 1, x_5 = 1$$

Che tradotto in linguaggio naturale significa: il primo individuo intervistato ha zero materie insufficienti, il secondo due materie insufficienti, ecc.

AL LAVORO!

PROVA AD UTILIZZARE DIVERSE CODIFICHE PER I TUOI DATI.
PROVA A CODIFICARE UNA VARIABILE QUALITATIVA CHE HA PIÙ DI DUE MODALITÀ CON NUMERI, HA SENSO LA SOMMA DEI DATI?



1.3 LE VARIABILI STATISTICHE

Le Variabili Statistiche (VS) sono le caratteristiche di interesse misurate sull'unità statistica, si indicano con una lettera latina, solitamente la x e sono chiamate variabili perché variano da individuo ad individuo. Le possiamo dividere in base alla tipologia di caratteristica che misuriamo sull'unità.

Le VS possono essere:

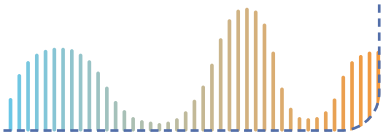
- **Qualitative**: se esprimono qualità non rappresentabili con numeri, se non per convenzione
  - ◊ **Non ordinate**: qualità il cui ordine è puramente convenzionale, per esempio

        - Sesso: Maschio o Femmina
        - Settore di Impiego: Agricoltura, Industria, Servizi
        - Paese di provenienza
    - **Ordinate**: qualità che sono ordinabili ma la cui distanza non è calcolabile
        - Massimo titolo di studio: primaria, media inferiore, media superiore, università, post università.
        - Livello di inquadramento lavorativo: Impiegato, Quadro, Dirigente.
        - Le preferenze: decisamente no, più no che sì, più sì che no, decisamente sì.
- **Quantitative**: se esprimono caratteristiche che possono essere rappresentate da numeri
    - **Discrete**: quantità che si misurano con i numeri interi 0, 1, 2, … ecc. Tipicamente i conteggi
        - Numero di materie insufficienti per studente
        - Numero di figli per famiglia
        - Numero di volte che si è cercato lavoro nell'ultimo mese per disoccupato
    - **Continue**: quantità che si esprimono con numeri decimali, tipicamente le misure metriche
        - Statura, espressa in metri di ogni studente
        - Velocità, espressa in km/h di tutti i veicoli che sono passati al km 32 dell'A1 nel 2022
        - Reddito, espresso in euro, di tutti i contribuenti della provincia di Modena nel 2021.

Ogni VS è suscettibile di assumere diversi valori chiamati **modalità**, se le VS sono qualitative i simboli che utilizziamo per descrivere le modalità saranno etichette di comodo (M per maschi e F per femmine) mentre se le VS sono quantitative useremo i numeri.

AL LAVORO!

INDIVIDUA DUE VARIABILI QUALITATIVE NON ORDINATE, DUE QUALITATIVE ORDINATE, DUE QUANTITATIVE DISCRETE E DUE QUANTITATIVE CONTINUE.



1.4 ORGANIZZAZIONE SIMBOLICA DEI DATI

I dati possono essere tanti e ci serve una notazione simbolica per indicarli. Le VS servono a questo scopo. Il numero di unità osservate può essere, 10, 20, 50 e indicheremo il numero di dati con la lettera n. Le etichette simboliche con la x (le VS) e useremo la lettera i per indicare simbolicamente la posizione della osservazione. La rappresentazione simbolica dei dati è dunque

$$\text{Dati} = (x_1, x_2, x_3, \ldots, x_i, \ldots, x_n)$$

e si legge: il primo dato osservato è x_1, il secondo x_2, il terzo x_3, …, l'i-esimo x_i, … e l'ultimo, l'ennesimo, x_n.

Esempio: $n = 5$, x il sesso (M=maschio; F=femmina),

$$\text{Dati} = (x_1 = M, x_2 = F, x_3 = F, x_4 = M, x_5 = M).$$

La codifica usata è M per i maschi e F per le femmine. Se avessimo codificato diversamente, per esempio 0 per i maschi e 1 per le femmine avremmo ottenuto

$$\text{Dati} = (x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1).$$

Con il vantaggio che sommando le x otteniamo il numero di femmine nei dati, infatti

$$x_2 + x_3 + x_4 + x_5 = 0 + 1 + 1 + 0 + 1 = 3,$$

sono 3 le femmine su 5.

Esempio: $n = 10$, x diploma (L= liceo, T= istituto tecnico, P=istituto professionale)

$$\text{Dati} = (x_1 = L, x_2 = L, x_3 = T, x_4 = P, x_5 = T, x_6 = P, x_7 = L, x_8 = T, x_9 = T, x_{10} = L).$$

Attenzione: codificare numericamente variabili qualitative che hanno più di due modalità può risultare complesso.

Esempio: $n = 8$, x reddito lordo mensile (espresso in migliaia di euro)

Dati $= (x_1 = 3.6,\ x_2 = 2.4,\ x_3 = 5.2,\ x_4 = 3.2,\ x_5 = 4.3,\ x_6 = 2.2, x_7 = 2.9, x_8 = 1.2)$.

Se su ogni individuo vogliamo rilevare due o più caratteristiche, per esempio il tipo di diploma x, il reddito mensile y e il numero libri letti all'anno z, è molto comodo mettere per riga l'individuo e per colonna le variabili, dando vita alla matrice dei dati per esempio

Num. progressivo	tipo di diploma x	reddito mensile y	numero libri z
1	$x_1 = L$	$y_1 = 3.5$	$z_1 = 6$
2	$x_2 = T$	$y_2 = 2.6$	$z_2 = 8$
3	$x_3 = T$	$y_3 = 4.1$	$z_3 = 0$
4	$x_4 = P$	$y_4 = 1.6$	$z_4 = 2$
...	...	...	...

AL LAVORO!

CREA ALMENO QUATTRO VARIABILI, DI UNA MATRICE DEI DATI.

1.5 ORDINAMENTO E CONTEGGIO

Se l'ordine di osservazione non è rilevante i dati si possono ordinare a piacimento. Se i dati sono quantitativi (ovvero numerici) si possono ordinare dal più piccolo al più grande, se per esempio i dati sono:

$$x_1 = 3.6,\ x_2 = 2.4,\ x_3 = 5.2,\ x_4 = 3.2,\ x_5 = 4.3,\ x_6 = 2.2, x_7 = 2.9, x_8 = 1.2$$

i dati ordinati saranno

$$x_{(1)} = 1.2,\ x_{(2)} = 2.2,\ x_{(3)} = 2.4,\ x_{(4)} = 2.9,\ x_{(5)} = 3.2,\ x_{(6)} = 3.6, x_{(7)} = 4.3, x_{(8)} = 5.2$$

dove $x_{(1)}$ indica il più piccolo dei dati ordinati, $x_{(2)}$ indica il più secondo piccolo dei dati ordinati, …, $x_{(8)}$ indica il più grande dei dati ordinati.

Se le etichette che identificano la codifica non hanno un ordine definito, ad esempio

$$x_1 = M,\ x_2 = F,\ x_3 = F,\ x_4 = M,\ x_5 = M$$

diventa

$$x_{(1)} = M,\ x_{(2)} = M,\ x_{(3)} = M,\ x_{(4)} = F,\ x_{(5)} = F$$

se abbiamo stabilito che M viene prima di F, ma

$$x_{(1)} = F,\ x_{(2)} = F,\ x_{(3)} = M,\ x_{(4)} = M,\ x_{(5)} = M$$

è ugualmente corretta.

1.6 DISTRIBUZIONI DI FREQUENZA

Una volta che i dati sono riordinati possiamo contare con più facilità le frequenze (il numero di volte) con cui una modalità compare. Definiamo la **frequenza assoluta** n_j, in numero di volte in cui compare la modalità j. Le frequenze assolute possono essere relativizzate dividendole per n, definiamo $f_j = n_j / n$ la **frequenza relativa** della modalità j. Infine definiamo $f_{j\%} = f_j \times 100$ le **frequenze percentuali**.

Le distribuzioni di frequenza sono tabelle che associano ad ogni modalità della variabile la frequenza (il numero di volte) con cui vengono osservate, nel nostro piccolo esempio con 5 dati di cui 3 maschi e 2 femmine la distribuzione è la seguente

Sesso	n_j	$f_{j\%} = n_j / n \times 100$
M	3	60%
F	2	40%
Totale	5	100%

Le n_j sono le frequenze e l'indice j serve a contare le modalità, in questo caso solo 2, $j = 1$, M, e $j = 2$ per F.

Più formalmente, se consideriamo una VS che un numero k di modalità espresse con le etichette, m_1, m_2, ..., m_j, ..., m_k. Una distribuzione di frequenza è la tabella

X	n_j	$f_{j\%}$
m_1	n_1	$f_{1\%}$
m_2	n_2	$f_{2\%}$
...	...	...
m_j	n_j	$f_{j\%}$
...	...	...
m_j	n_j	$f_{j\%}$
Totale	n	100%

e si leggerà: n_1 individui presentano la modalità m_1, n_2 individui presentano la modalità m_2, ..., n_j individui presentano la modalità m_j, e così via fino all'ultima (k).

Se i dati sono ordinati possiamo costruire una nuova frequenza ottenuta partendo dalle frequenze percentuali, chiamata frequenza cumulata. La frequenza cumulata è indicata con F ed è ottenuta sommando le f.

$$F_1 = f_1$$

$$F_2 = f_1 + f_2$$

$$F_3 = f_1 + f_2 + f_3$$

...

A titolo di esempio qui osserviamo le mie statistiche sulla qualità dell'insegnamento. Sono espresse, nella tabella qui sotto, le risposte alla domanda: *Il docente stimola/motiva l'interesse verso la disciplina?*

	Decisamente no	Più no che sì	Più sì che no	Decisamente sì	
Frequenza	1	10	69	66	146
Percentuale	0.68%	6.85%	47.26%	45.21%	100.00%
Percentuale cumulata	0.68%	7.53%	54.79%	100.00%	
Retro cumulata	100.00%	99.32%	92.47%	45.21%	

Su 146 studenti che hanno risposto, uno risponde "decisamente no" (lo 0.68%), 10 "più no che sì" (il 6.85%), 69 "più sì che no" (47.26%) e 66 "decisamente sì". Siccome le modalità sono ordinate possiamo cumulare le frequenze percentuali. Le frequenze cumulate si interpretano nel seguente modo: lo 0.68% risponde "decisamente no", il 7.53% risponde *non meglio di* "più no che sì", il 54.79% risponde *non meglio di* "più sì che no" e, ovviamente il 100% del collettivo risponde *non meglio di* "decisamente sì". Se cumuliamo partendo dall'ultima modalità otteniamo le frequenze retrocumulate. Il 100% del collettivo ha risposto *non peggio di* "decisamente no", il 99.32% ha risposto *non peggio di* "più no che sì", il 92.37% del collettivo ha risposto più sì che no.

Se i dati sono espressi in forma numerica con tante modalità, conviene raccoglierli in classi. A titolo esemplificativo abbiamo raccolto i dati delle dichiarazioni dei redditi del 2021, scaricabili gratuitamente dal sito dell'agenzia dell'entrate all'indirizzo https://www1.finanze.gov.it/finanze/analisi_stat/public/index.php?opendata=yes. Nella tabella 1 mostriamo una elaborazione dei dati dell'Emilia-Romagna. I dati sono espressi in migliaia di euro.

	Classi di reddito			n_j	$f_{j\%}$	$F_{j\%}$
	meno di		5	461,833	13.65%	13.65%
Tra	5	e	10	365,906	10.81%	24.46%
Tra	10	e	15	417,875	12.35%	36.81%
Tra	15	e	20	503,489	14.88%	51.69%
Tra	20	e	25	509,982	15.07%	66.77%
Tra	25	e	30	375,267	11.09%	77.86%
Tra	30	e	40	389,299	11.51%	89.36%
Tra	40	e	50	142,384	4.21%	93.57%
	più di		50	217,594	6.43%	100.00%
	Totale			3,383,629	100.00%	

Tabella 1. Distribuzione dei redditi lordi dichiarati all'Agenzia delle Entrate nell'anno 2021 in Emilia-Romagna. Le classi sono espresse in migliaia di euro. (elaborazione su dati dell'Agenzia dell'Entrate)

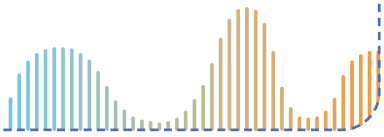
L'analisi della tabella 1 ci dice che in Emilia-Romagna sono state registrate, nell'anno 2021, 3,383,629 dichiarazioni dei redditi relative ad altrettante persone fisiche. 461,833 di esse ha dichiarato di avere un reddito lordo annuo inferiore a 5 mila euro e rappresentano il del totale; 365,906 dichiarano un reddito lordo annuo compreso tra 5mila e 10mila euro l'anno e rappresentano del totale; e così via fino all'ultima classe.

La colonna delle percentuali cumulate è di grande interesse perché ci consente di osservare alcuni aspetti della distribuzione dei dati. Le percentuali cumulate si leggono: il 13.65% delle dichiarazioni è minore di 5 mila, il 24.46% delle dichiarazioni è minore di 10 mila, il 36.81% delle dichiarazioni è minore di 20 mila, ecc. ecc.

Osserviamo per esempio che circa un quarto (24.46%) delle dichiarazioni si attestano sotto i 10 mila e di conseguenza il 75% circa delle dichiarazioni sono superiori a 10 mila. Circa la metà delle dichiarazioni (51.69%) è inferiore a 20 mila e di conseguenza la rimanente metà dichiara più di 20 mila. Osserviamo infine che circa il 75% del collettivo (il 77.86%) ha un reddito dichiarato inferiore a 30mila e quindi circa il 25% ha un reddito dichiarato superiore a 30mila.

AL LAVORO!

SCARICA I DATI IN FORMATO ELETTRONICO DALLA AGENZIA DELLE ENTRATE E PROVA AD OSSERVARE I DATI CON UN FOGLIO ELETTRONICO, COME EXCEL O I FOGLI DI GOOGLE.



1.7 DISTRIBUZIONI DI DUE VARIABILI

Se dobbiamo osservare come due variabili variano insieme possiamo costruire delle distribuzioni di frequenza doppie, ovvero distribuzioni che si possono leggere per riga per una variabile e per colonna per l'altra. Se il nostro obiettivo è rappresentare due variabili espressa in *k* modalità e, espressa in *h* modalità. Dobbiamo costruire la tabella delle frequenze **congiunte**, anche nota come **tabella di contingenza**, come evidenziato in tabella 2.

	y_1	y_2	…	y_j	…	y_h	Tot
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1h}	n_1
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2h}	n_2
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ih}	n_j
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kh}	n_h
Tot	n_1	n_2	…	n_j	…	n_h	n

Tabella 2. Rappresentazione simbolica di una tavola di contingenza

Dove n_j è il numero di individui che mostrano la caratteristica x_i e la caratteristica y_j. I totali si ottengono sommando per riga e per colonna

$n_i = n_{i1} + n_{i2} + \ldots + n_{ih}$

è la somma della *i*-esima riga: il totale degli individui che presenta la caratteristica x_i, indipendentemente da *y*, mentre

$n_j = n_{1j} + n_{2j} + \ldots + n_{kj}$

è la somma della *j*-esima colonna: il totale degli individui che presenta la caratteristica y_j, indipendentemente da *x*.

A partire da questa tabella possiamo costruire diverse altre frequenze, quali le frequenze relative congiunte $f_{ij} = \frac{n_{ij}}{n} \times 100$, le frequenze percentuali condizionate alle righe $f_{j|i} = \frac{n_{ij}}{n_i} \times 100$ e quelle condizionate alle colonne $f_{i|j} = \frac{n_{ij}}{n_j} \times 100$.

Per esempio, abbiamo aggregato di dati delle dichiarazioni dei redditi per le regioni del nord, del centro e del sud (isole comprese), i dati sono rappresentati in tabella 3.

	Classi di reddito			Nord	Centro	Sud	
	meno di		5	2,468,819	2,126,518	2,557,519	7,152,856
tra	5	e	10	1,959,665	1,681,346	2,156,579	5,797,590
tra	10	e	15	2,068,315	1,669,346	1,650,654	5,388,315
tra	15	e	20	2,464,503	1,777,301	1,328,206	5,570,010
tra	20	e	25	2,527,805	1,681,383	1,110,003	5,319,191
tra	25	e	30	1,839,872	1,246,474	836,738	3,923,084
tra	30	e	40	1,934,791	1,366,744	918,39	4,219,925
tra	40	e	50	717,248	513,76	290,897	1,521,905
	più di		50	1,130,767	778,134	374,005	2,282,906
				17,111,785	12,841,006	11,222,991	41,175,782

Tabella 3. Tabella di contingenza delle frequenze assolute che mostra le classi di reddito (espresse in migliaia di euro) sulle righe e la posizione geografica sulle colonne. (elaborazione su dati dell'Agenzia dell'Entrate)

In tutta Italia sono 7,152,856 i contribuenti che dichiarano meno di 5 mila euro l'anno, 2,468,819 al nord, 2,126,518 al centro e 2,557,519 al sud, mentre i contribuenti che dichiarano più di 50 mila euro l'anno sono 2,282,906, 1,130,76 al nord, 778,134 al centro e 374,005 al sud. 17,111,785 sono le dichiarazioni al Nord, 12,841,006 al centro e 11,222,991 al sud, per un totale di 41,175,782 di contribuenti.

Se dividiamo tutte le frequenze per 41,175,782 e moltiplichiamo per 100, otteniamo le frequenze percentuali congiunte che leggiamo in tabella 4.

	Classi di reddito			Nord	Centro	Sud	
	meno di		5	6.00%	5.16%	6.21%	17.37%
tra	5	e	10	4.76%	4.08%	5.24%	14.08%
tra	10	e	15	5.02%	4.05%	4.01%	13.09%
tra	15	e	20	5.99%	4.32%	3.23%	13.53%
tra	20	e	25	6.14%	4.08%	2.70%	12.92%
tra	25	e	30	4.47%	3.03%	2.03%	9.53%
tra	30	e	40	4.70%	3.32%	2.23%	10.25%
tra	40	e	50	1.74%	1.25%	0.71%	3.70%
	più di		50	2.75%	1.89%	0.91%	5.54%
				41.56%	31.19%	27.26%	100.00%

Tabella 4. Tabella di contingenza delle frequenze percentuali che mostra le classi di reddito (espresse in migliaia di euro) sulle righe e la posizione geografica sulle colonne. Si ottiene a dividendo le frequenze assolute congiunte per il numero totale di dichiarazioni e moltiplicando per 100. (elaborazione su dati dell'Agenzia dell'Entrate)

Il 6% delle dichiarazioni delle dichiarazioni italiane è sotto ai 5 mila euro ed è al nord Italia, il 5.16% è al centro e il 6.21% al sud, per un totale del 17.37% di dichiarazioni sotto ai 5 mila euro. Mentre è interessante osservare che le dichiarazioni sopra i 50 mila euro sono il 5.54% delle dichiarazioni nazionali di cui il 2.75% al nord, l'1.89% al centro e lo 0.91% al sud.

Se dividiamo le frequenze congiunte per i totali di riga otteniamo la tabella delle **frequenze condizionate** alla classe di reddito, che ci dice com'è ripartita ogni classe di reddito tra nord, centro e sud, come mostrato in tabella 5.

	Classi di reddito			Nord	Centro	Sud	
	meno di		5	34.52%	29.73%	35.76%	100.00%
tra	5	e	10	33.80%	29.00%	37.20%	100.00%
tra	10	e	15	38.39%	30.98%	30.63%	100.00%
tra	15	e	20	44.25%	31.91%	23.85%	100.00%
tra	20	e	25	47.52%	31.61%	20.87%	100.00%
Tra	25	e	30	46.90%	31.77%	21.33%	100.00%
Tra	30	e	40	45.85%	32.39%	21.76%	100.00%
Tra	40	e	50	47.13%	33.76%	19.11%	100.00%
	più di		50	49.53%	34.09%	16.38%	100.00%
				41.56%	31.19%	27.26%	100.00%

Tabella 5. Tabella delle frequenze condizionate alle classi di reddito. Si ottiene a dividendo le frequenze congiunte per il numero totale di riga. (elaborazione su dati dell'Agenzia dell'Entrate)

Il 34.52% di chi dichiara meno di 5mila euro risiede al nord, il 29.73% al centro e il 35.76% al sud. Mentre tra chi dichiara più di 50 mila è quasi per il 50% (49.53%) dei casi residente al nord, il 34% al centro e solo il 16.38% al sud.

Dividere per i totali di colonna offre un'altra rappresentazione dei dati ancora: la tabella delle **frequenze condizionate** all'area geografica, ci dice come in ogni regione sia distribuito il reddito, come mostrato in tabella 6.

	Classi di reddito			Nord	Centro	Sud	
	meno di		5	14.43%	16.56%	22.79%	17.37%
tra	5	e	10	11.45%	13.09%	19.22%	14.08%
tra	10	e	15	12.09%	13.00%	14.71%	13.09%
tra	15	e	20	14.40%	13.84%	11.83%	13.53%
tra	20	e	25	14.77%	13.09%	9.89%	12.92%
tra	25	e	30	10.75%	9.71%	7.46%	9.53%
tra	30	e	40	11.31%	10.64%	8.18%	10.25%
tra	40	e	50	4.19%	4.00%	2.59%	3.70%
	più di		50	6.61%	6.06%	3.33%	5.54%
				100.00%	100.00%	100.00%	100.00%

Tabella 6. Tabella delle frequenze condizionate alla posizione geografica. Si ottiene a dividendo le frequenze congiunte per il numero totale di colonna. (elaborazione su dati dell'Agenzia dell'Entrate)

Nella prima riga si legge che dichiarano meno di 5 mila euro: il 14.43% dei contribuenti del nord, il 16.56% dei contribuenti del centro e il 22.79% di quelli del sud, mentre è il 17.37% dei contribuenti totali italiani.

AL LAVORO!

RILEVA SUI TUOI COMPAGNI IL VOTO IN ITALIANO E IL VOTO IN MATEMATICA E CREA UNA TABELLA DI CONTINGENZA DOVE PER RIGA METTI IL VOTO DI ITALIANO E PER COLONNA QUELLO DI MATEMATICA. COSTRUISCI LE RELATIVE FREQUENZE CONDIZIONATE AL VOTO DI ITALIANO E A QUELLO DI MATEMATICA.

1.8 RAPPRESENTAZIONI GRAFICHE

Confrontare le distribuzioni dalla tabella può essere molto difficoltoso, la rappresentazione grafica delle tabelle è di aiuto. Se i dati sono qualitativi si possono costruire **grafici a barre**, dove si associa ad ogni modalità un rettangolo di altezza pari alle frequenze, oppure **grafici a torta**, dove si divide una circonferenza in spicchi di ampiezza proporzionale alle frequenze. Se per esempio volessi rappresentare il mio mini-collettivo di 3 maschi e 2 femmine potrei optare tra uno dei due grafici di figura 1.

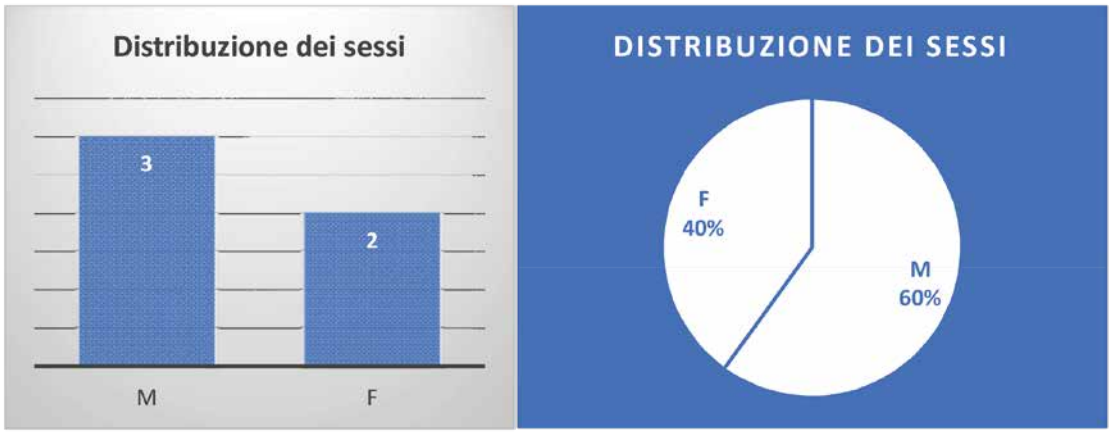
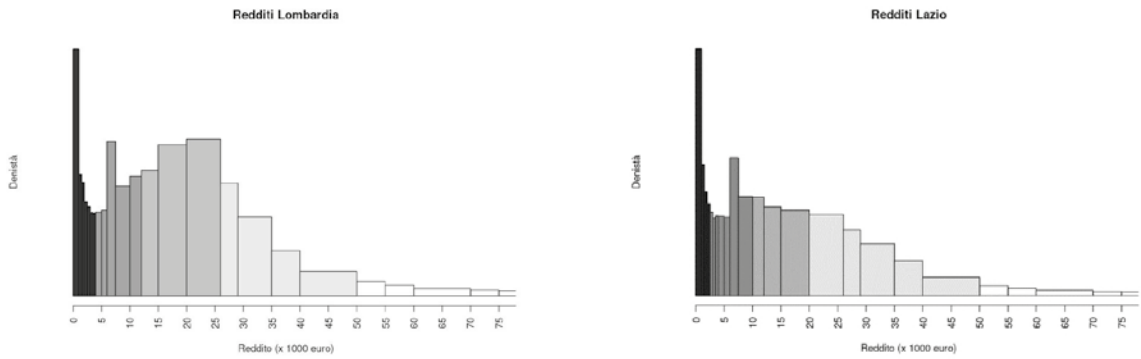


Figura 1 Distribuzione dei sessi rappresentata con un diagramma a barre a sinistra e con una torta a destra.

Se i dati sono raccolti in classi di ampiezze diverse è necessario aggiustare le altezze dei i rettangoli con l'ampiezza della classe dando luogo agli **istogrammi di densità**. L'obiettivo è di rappresentare rettangoli contigui, di larghezza pari all'ampiezza delle classi e di altezza ottenuta in modo tale che l'area sottesa sia pari alla frequenza, ovvero l'altezza dei rettangoli è data dalla formula

$$h_j = Const. \frac{f_j}{b_j}$$

Dove, è una constante arbitraria, se otteniamo l'istogramma delle frequenze relative, se quello delle frequenze percentuali e, infine, se quello delle frequenze assolute. A titolo esemplificativo rappresentiamo in figura 2. l'istogramma delle densità relative, della regione Lombardia e della regione Sicilia, con la distribuzione in classi completa rilasciata dall'Agenzia delle Entrate, che è molto più raffinata di quella esemplificata negli esempi proposti in precedenza.



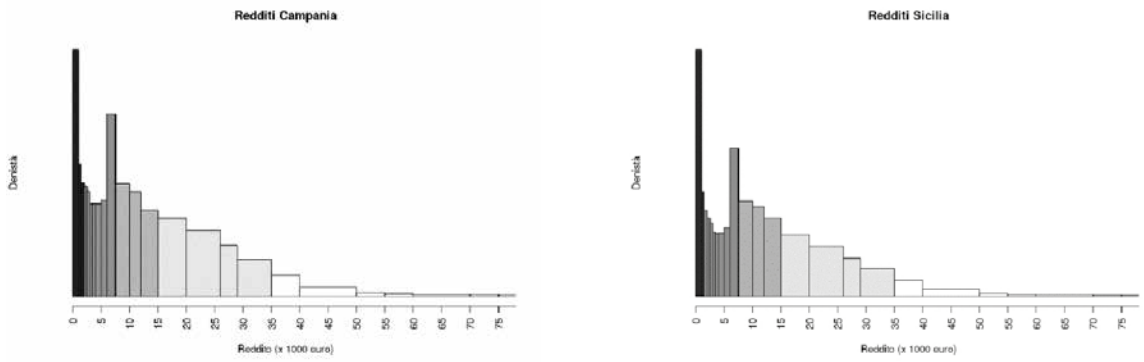


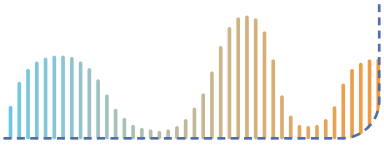
Figura 2. Istogramma di densità delle regioni Lombardia, Lazio, Campania e Sicilia a confronto. (elaborazione dati dell'Agenzia delle Entrate).

In nero è colorato il primo quinto (il 20%) delle dichiarazioni più basse, in grigio scuro il secondo quinto, in grigio il terzo quinto, in grigio chiaro il quarto quinto e in bianco il l'ultimo quinto delle dichiarazioni più alte, regione per regione. È interessante osservare la differenza tra la distribuzione dei redditi in Lombardia e quella delle altre regioni. Il reddito in Lombardia è decisamente spostato su redditi più elevati, raggiungendo un picco tra i 20 e i 25 mila euro annui. Il 50% dei redditi lombardi sono inferiori ai 27.5 mila euro circa, mentre il 20% dei più alti è superiore ai 50 mila euro annui. In Lazio il 50% dei redditi è inferiore ai 20 mila euro, mentre in Campania e in Sicilia il 50% dei redditi è inferiore ai 15 mila euro.

AL LAVORO!

QUALI ALTRE CONCLUSIONI POSSIAMO TRARRE SULL'ANALISI DEI QUATTRO ISTOGRAMMI IN FIGURA 2?

COSTRUISCI GLI ISTOGRAMMI DI DENSITÀ RELATIVA, PERCENTUALE E ASSOLUTA DEI DATI NORD, CENTRO E SUD PROPOSTI IN TABELLA 3. QUALE DIFFERENZA VEDI TRA I TRE? RACCOGLI DATI QUANTITATIVI, METTILI CLASSI E RAPPRESENTA L'ISTOGRAMMA. COSA CAMBIA CAMBIANDO LA DEFINIZIONE DELLE CLASSI?



CAPITOLO 2. LA STATISTICA PER I DATI STORICI

Nella matrice dei dati le unità statistiche sono costituite in genere da individui, oggetti, aziende, ecc.
Una seconda categoria di matrici è quella in cui i valori delle variabili rilevate sono riferiti ad unità temporali.

Le unità temporali possono essere:

- «tempi» ⇒ ogni valore è riferito ad un istante (ad esempio, popolazione in Italia al 31/12 di ogni anno)
- «intervalli temporali» ⇒ i valori riguardano un arco di tempo (ad esempio, numero di nascite mensili in Italia nell'ultimo anno)

Si dice **serie storica (o serie temporale) semplice** una successione di valori d'una variabile quantitativa, riferiti a tempi o intervalli temporali. Quando le variabili sono più di una, si parla di **serie storica multipla.**

Esempio: serie storica Istat sulla pratica sportiva e sedentarietà
Nella tabella viene riportato il numero di persone (in migliaia) in Italia, di 3 anni e più, che svolge o non svolge la pratica sportiva.

	PRATICANO SPORT			
ANNI	IN MODO CONTINUATIVO	IN MODO SALTUARIO	SOLO QUALCHE VOLTA	NON PRATICANO SPORT
2011	12717	5878	16017	23074
2012	12743	5395	16992	22790
2013	12602	5364	16341	24156
2014	13582	5069	16540	23518
2015	14013	5603	15607	23524
2016	14792	5693	15108	23085
2017	14607	5365	16273	22426
2018	15107	5623	16748	21087
2019	15605	4905	17234	20895
2020	15837	5559	16419	20583

21 milioni e 396 mila persone di 3 anni e più nel 2020 praticavano uno o più sport nel loro tempo libero, di questi 15 milioni 837 mila praticavano sport con assiduità e 5 milioni 559 mila saltuariamente. 20 milioni 583 mila persone di 3 anni e più nel 2020 dichiaravano di NON praticare sport né attività sportiva e di essere sostanzialmente sedentari. Tra il 2013 ed il 2020 si rilevano oltre 3 milioni e mezzo di sedentari in meno nella popolazione italiana.

Sarebbe interessante capire se nel corso degli anni il numero di sportivi è aumentato o diminuito, se vi è stato un trend di crescita o di decrescita e quale è stata la velocità di crescita o di decrescita negli ultimi anni (velocità che, presumibilmente, si rifletterà nel numero di sportivi e sedentari stimati per gli anni successivi al 2020).

Sul sito del CONI, nella sezione «I numeri dello Sport» si legge: «Le statistiche dell'ISTAT rilevano un trend in crescita tra i praticanti che svolgono attività sportiva in modo continuativo nel proprio tempo libero: +8 punti percentuali dal 2001 al 2020. Nel 2020 le persone di 3 anni e più che praticavano sport con continuità erano il 27,1% della popolazione. Nell'ultimo periodo la velocità di crescita è stata maggiore: si è passati dal 21,5% del 2013 al 27,1% del 2020, recuperando 5,6 punti percentuali. Contestualmente, la sedentarietà è diminuita di 6 punti percentuali scendendo dal 41,2% rilevato nel 2013 al 35,2% nel 2020»

Come si arriva a queste considerazioni? Quali sono le analisi statistiche che studiano i fenomeni rilevati su scala temporale? Cerchiamo di rispondere a queste domande nei prossimi paragrafi.

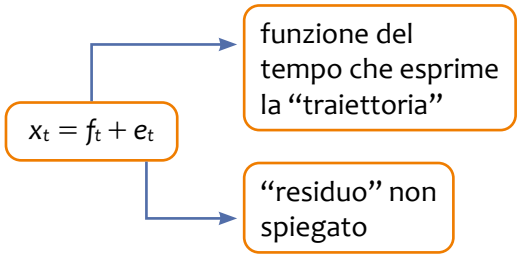
2.1 LA STIMA DEL TREND

- Per sapere se al trascorrere del tempo il fenomeno ha presentato una tendenza, detta trend, all'aumento o alla diminuzione (o se è rimasto allo stesso livello) vi sono analisi statistiche mirate alla stima del trend
- Il trend può interpretarsi come la "traiettoria ideale" che la variabile avrebbe percorso nella sua evoluzione temporale **in assenza** di perturbazioni dovute a **fattori erratici** o **accidentali**, a **componenti stagionali** (ad esempio, se analizzo la serie storica delle vendite di gelato, sicuramente avrò un calo in autunno e inverno ed un aumento in primavera ed estate) o **cicliche** (ad esempio, il numero di chi pratica attività fisica saltuariamente aumenta dopo le ferie)
- La valutazione del trend è riferita all'intervallo di tempo considerato ed è su questo che si vuole individuare una tendenza di fondo. Nel caso d'una serie storica con dati annuali si può parlare di trend con riferimento ad un arco temporale di almeno 10 anni; per una serie storica con dati giornalieri (ed esempio, le quotazioni in Borsa d'un titolo azionario) si può invece individuare anche la tendenza nell'ultimo mese o nell'ultimo trimestre

Per stimare i valori del trend i 2 metodi principali sono:

- **Il metodo delle medie mobili (m.m.)**
- **Il metodo delle funzioni interpolanti**

In entrambi i casi si assume la relazione:



- Con le funzioni interpolanti la stima di f_t si ottiene scegliendo una funzione analitica (retta, parabola, …) da adattare ai valori osservati della serie
- Nelle medie mobili si assume che i residui presentino una successione di valori positivi e negativi (risultino superiori e inferiori rispetto alla traiettoria regolare) e che quindi una media d'un sufficiente numero di valori consecutivi della serie tenda a compensare gli effetti erratici, fornendo una stima del trend al tempo *t*.

Il metodo delle medie mobili

Si dice media mobile di ordine *s* la media aritmetica di *s* termini consecutivi della serie. Si assume ***s* dispari**, e si "centra" il valore di ogni media nel tempo di mezzo tra quelli considerati. Ad esempio, il valore al tempo 2 della m.m. di 3 termini è la media dei valori al tempo 1, 2 e 3. La m.m. di 5 termini al tempo 3 è la media dei valori al tempo 1, 2, 3, 4, 5. Il valore al tempo 4 della m.m. di 3 termini è la media dei valori al tempo 3, 4 e 5. La tabella riporta le medie mobili di 3 e 5 termini del numero di coloro che praticano sport con continuità. A titolo esemplificativo riportiamo il calcolo dettagliato di alcuni valori:

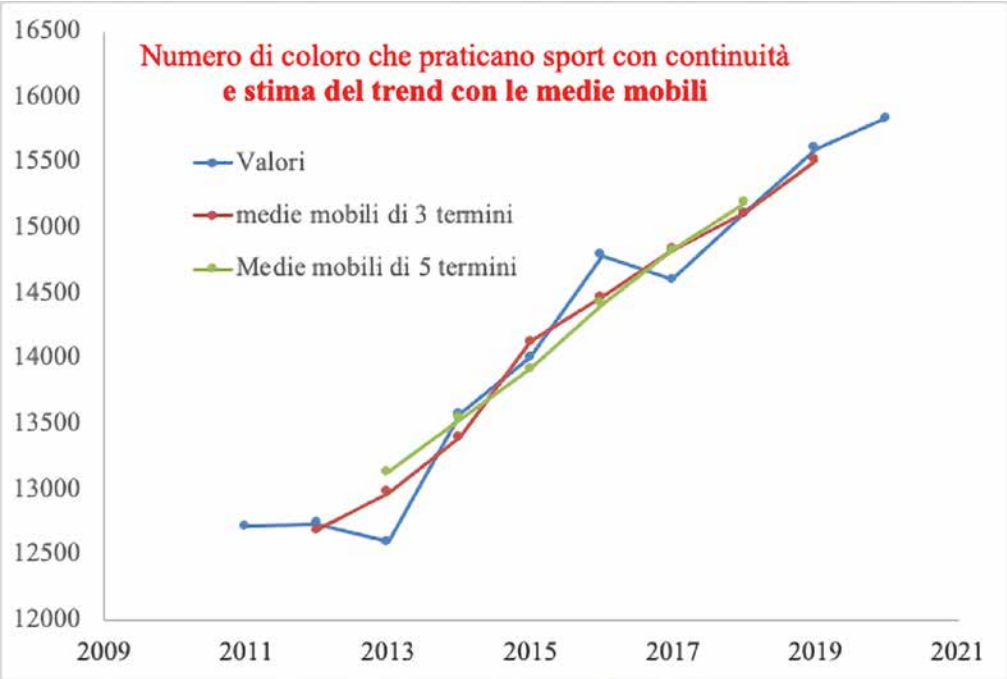
- m.m. di 3 termioni al 2013: 12976 = (12743+12602+13582) / 3
- m.m. di 5 termini al 2015: 13919 = (12602+13582+14013+14792+14607) / 5

È evidente come la stima del primo e ultimo tempo con le medie mobili di ordine 3 non possa essere effettuata (mancano i termini da mediare) cosi come la stima del primo, secondo, ultimo e penultimo tempo con le medie mobili di ordine 5.

- maggiore è l'ordine della media mobile,
- maggiore il numero di stime che non possono essere fatte!

ANNI	tempi	Praticano sport in modo continuativo	Medie Mobili di 3 termini	Medie Mobili di 5 termini
2011	1	12717		
2012	2	12743	12687	
2013	3	12602	12976	13131
2014	4	13582	13399	13546
2015	5	14013	14129	13919
2016	6	14792	14471	14420
2017	7	14607	14835	14825
2018	8	15107	15106	15190
2019	9	15605	15516	
2020	10	15837		

Il grafico riporta l'andamento dei valori e l'andamento del trend stimato sia con le medie mobili di ordine 3 sia con le medie mobili di ordine 5.



Si nota come l'andamento dei valori originari sia oscillatorio (in alcuni anni crescente, in altri decrescente) ma il trend sia DECISAMENTE crescente con una velocità di crescita quasi costante nelle medie mobili di 5 termini e leggermente diversa da anno in anno con quelle di 3 termini.
In sintesi, si vede come l'impiego delle m.m. consenta di eliminare, o ridurre fortemente, le oscillazioni erratiche della serie. L'effetto di "lisciamento" (*smoothing*) è tanto più efficace quanto maggiore è la lunghezza delle m.m. utilizzate, per una migliore compensazione tra valori in eccesso e valori in difetto rispetto al trend.

Vantaggi e svantaggi del metodo delle medie mobili

Svantaggi:

- Non permette di disporre di informazioni sull'andamento tendenziale più recente, spesso il più importante per poter fare previsioni.
- Maggiore è la lunghezza delle m.m., maggiore è la perdita di valori stimati.

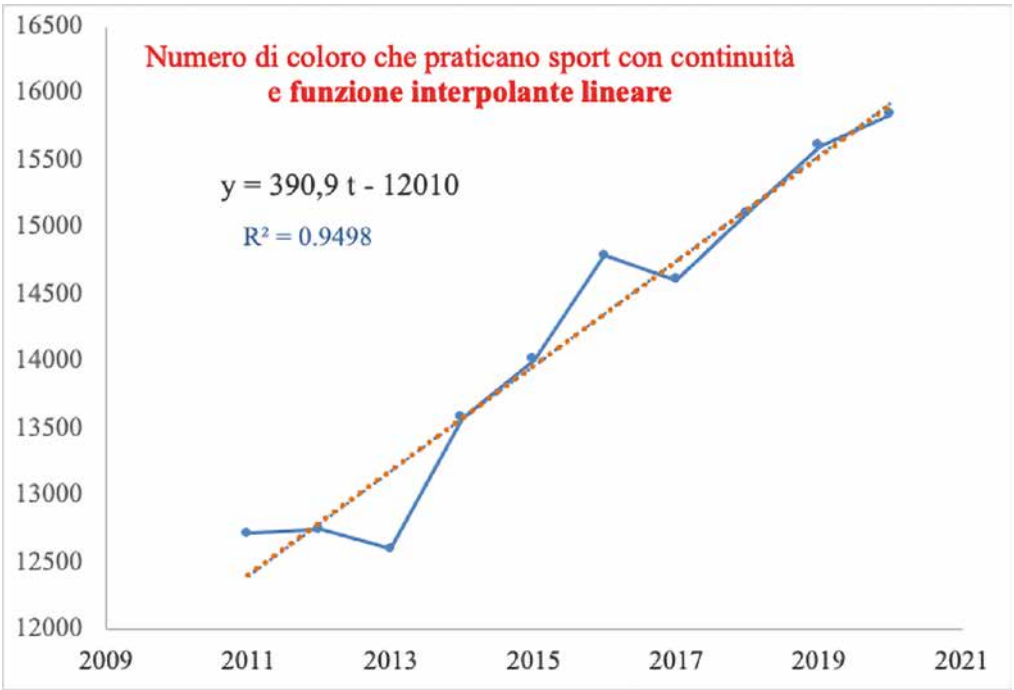
Vantaggi:

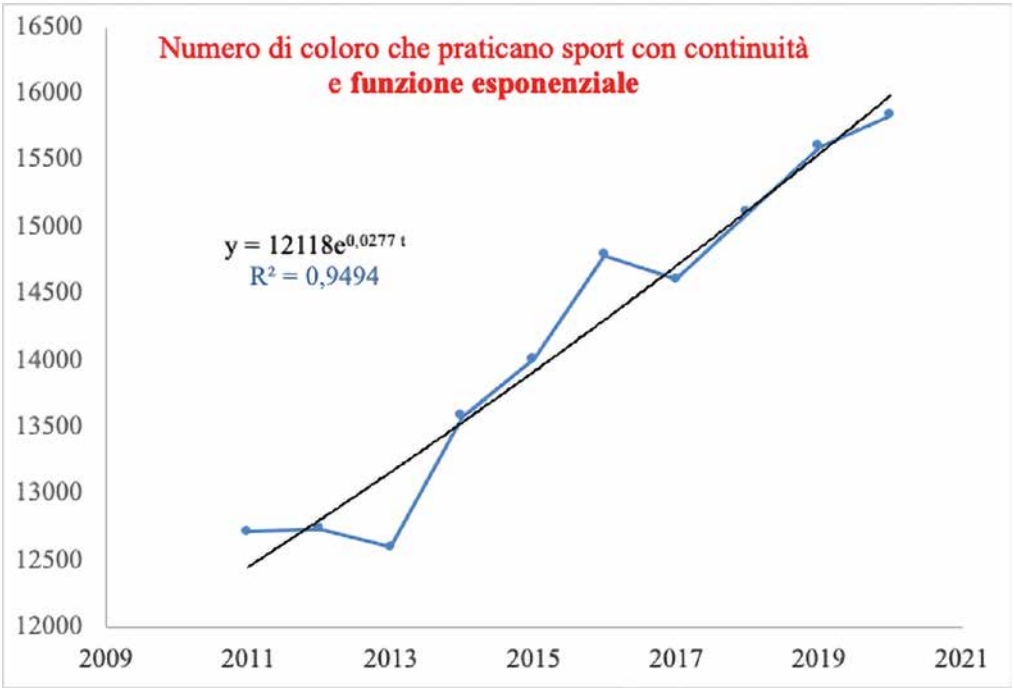
- Ha il pregio d'una estrema semplicità.
- Consente una stima dei valori del trend senza richiedere una scelta a priori del tipo di funzione atto a descrivere la tendenza di fondo.
- È più flessibile di altri metodi ed è in grado di adeguarsi meglio agli svariati tipi di evoluzione temporale che possono presentarsi nei differenti fenomeni.

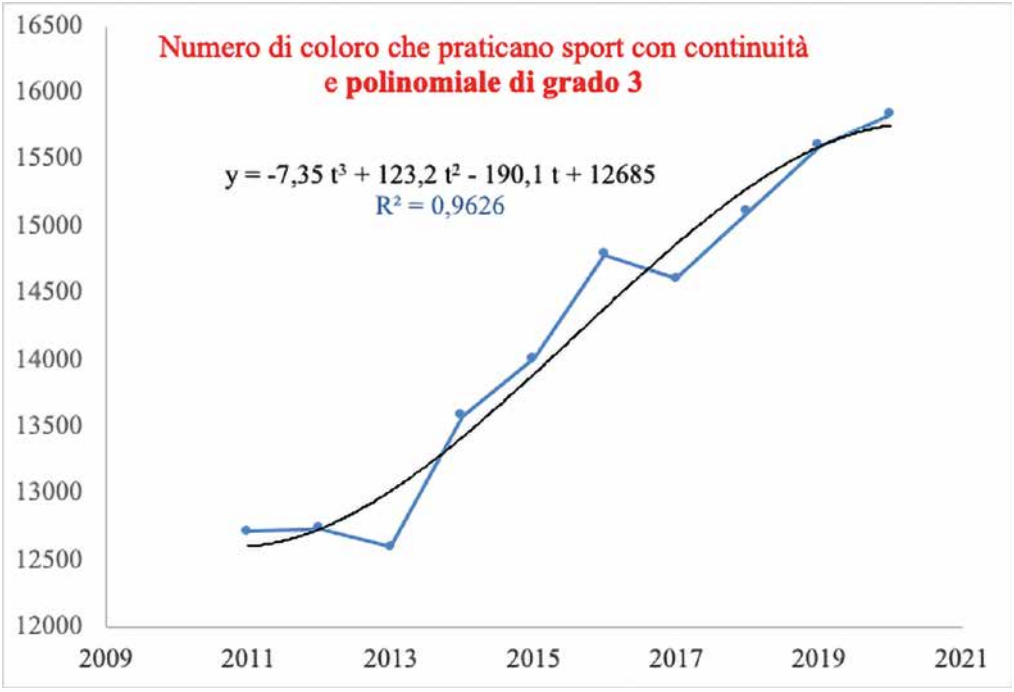
Il metodo delle funzioni interpolanti

Con le funzioni interpolanti si sceglie "a priori" una funzione analitica come ad esempio una retta, una parabola o una funzione logaritmica, da "adattare" ai dati. Il grafico dei valori può suggerire il tipo di funzione. Ad esempio, si nota dal grafico precedente come il numero di coloro che praticano sport in modo continuativo, dal 2010 al 2020, abbia un andamento nel tempo vicino alla linearità. I grafici successivi riportano alcune funzioni interpolanti stimate in base ai dati. Nei grafici vengono riportate le equazioni delle funzioni interpolanti con i parametri stimati e il valore dell'indice R^2

- Il metodo di stima dei parametri si chiama **"metodo dei minimi quadrati"** poiché minimizza la somma dei residui al quadrato.
- I valori della variabile indipendente sono i tempi (1, 2, …) come quelli inseriti in tabella
- La bontà di adattamento (e quindi l'affidabilità della stima del trend e delle previsioni che si possono fare utilizzando la funzione interpolante) è misurata da un **indice chiamato R^2.**
- R^2 varia da 0 ad 1 e misura la quota di variabilità "spiegata" dalla funzione interpolante. Maggiore è la quota di variabilità spiegata, migliore la bontà di adattamento. Se la funzione interpolante passasse esattamente fra tutti i punti osservati, R^2 sarebbe uguale a 1 (la funzione interpolante spiegherebbe il 100% della variabilità dei valori).





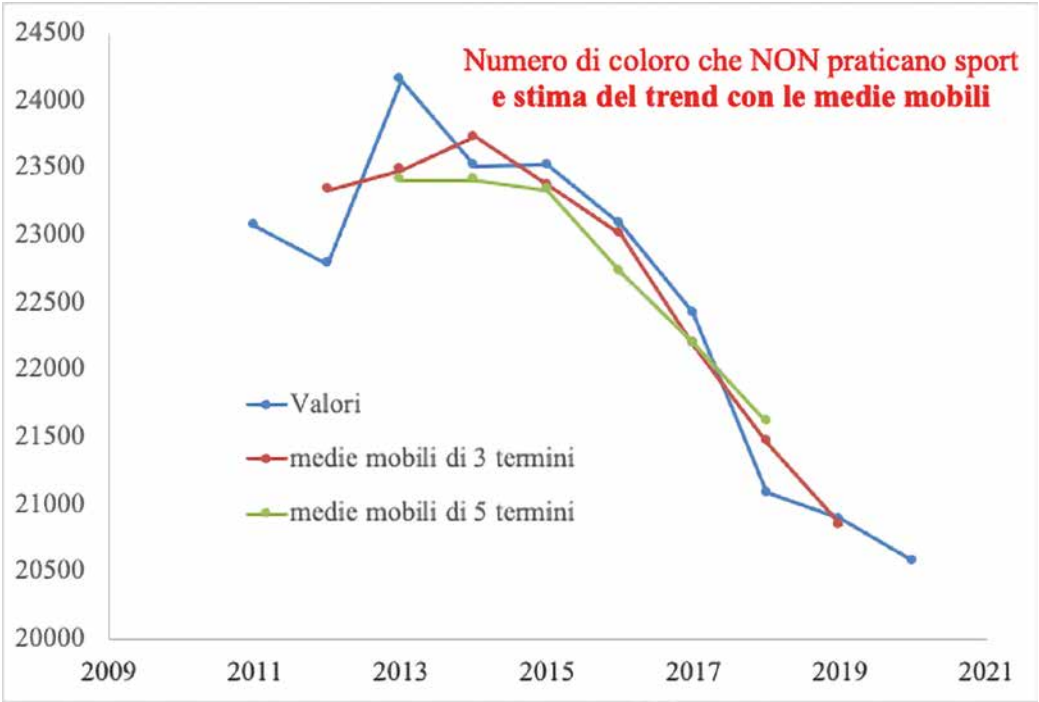


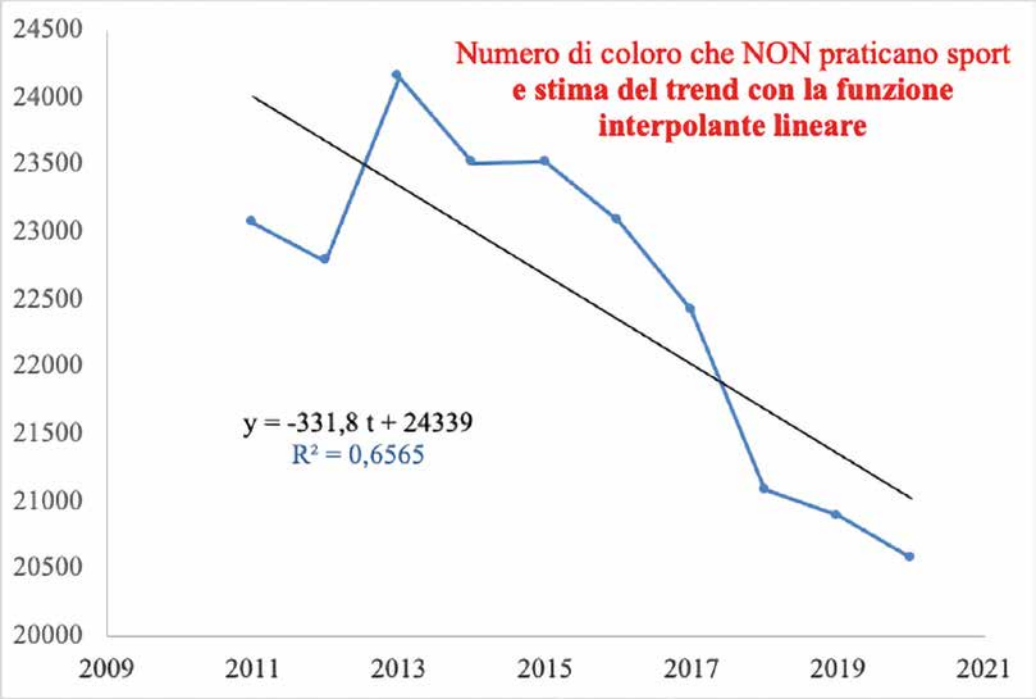
Si nota come tutte e tre le funzioni interpolanti stimate abbiano un ottimo adattamento. La funzione polinomilae di grado 3 ha l'indice R^2 maggiore ed è quindi quella con il migliore adattamento. Tuttavia, anche se non con l'indice R^2 maggiore, la funzione lineare è in generale preferibile per la sua semplicità e per la possibilità di interpretare i parametri. La pendenza è la stima dell'incremento (o decremento, se negativa) medio per unità di tempo. L'intercetta è il valore stimato al tempo precedente il primo rilevato. Dall'equazione della funzione lineare ricaviamo che il numero di sportivi stimato al 2010 è di 12010 unità e, in media, ogni anno il numero di coloro che praticano sport con continuità aumenta di quasi 391 unità. Sulla base di queste nozioni possiamo fare previsioni per stimare, ad esempio, il numero di coloro che praticheranno sport nel 2023 o nel 2024.

- Le previsioni per tempi successivi all'ultimo rilevato sono affidabili se e **solo se** la bontà di adattamento della funzione interpolante è prossima a 1 **e, congiuntamente,** la previsione non è per un tempo troppo distante dall'ultimo. Quest'ultima condizione esiste perché è plausibile che la forma del trend possa cambiare nel tempo. Ad esempio, non è detto che dal 2024 in poi il trend del numero di sportivi vari linearmente come nel decennio precedente.

AL LAVORO!

I GRAFICI SEGUENTI RIPORTANO LA SERIE DEL NUMERO DI SEDENTARI E LE STIME DEL TREND CON LE MEDIE MOBILI DI ORDINE 3 E 5 E LA FUNZIONE INTERPOLANTE LINEARE.
COSA POSSIAMO DIRE SULL'ANDAMENTO DI COLORO CHE NON FANNO SPORT?
CHE NUMERI CI ASPETTIAMO PER IL 2023 E IL 2024?





Suggerimento: a parte un aumento annuale del numero di sedentari dal 2012 al 2013 e dal 2014 al 2015, si nota un trend decisamente decrescente, più marcato a partire dal 2014 in poi. Si stima che il numero di sedentari diminuisca in media ogni anno di quasi 332 unità. Nel 2023 ci possiamo aspettare circa 20023 sedentari. La previsione è attendibile perché il tempo non è troppo distante dal 2020 (ultimo anno rilevato per la stima del trend) e l'adattamento della funzione interpolante lineare è discreto (R^2=0,65). Per il 2024 la previsione inizia a diventare poco attendibile. Non è plausibile pensare che il numero di sedentari possa continuare a dimunuire ogni anno di 332 unità.

2.2 LE VARIAZIONI ASSOLUTE E PERCENTUALI E I NUMERI INDICE

Nell'analisi delle serie storiche è interessante valutare come sia variato il fenomeno nel tempo e comparare la dinamica di diversi fenomeni. Le variazioni possono essere calcolate rispetto ad un tempo prefissato (**detto base**) che di solito è il primo rilevato, oppure di volta in volta rispetto al tempo precedente. Ad esempio, le **variazioni assolute** del numero di coloro che fanno sport dal 2011 al 2020 (e quindi nel 2020 rispetto al tempo base 2011) sono:

15837 – 12717 = +3120 migliaia di persone che praticano sport in modo continuativo
5559 – 5878 = –319 migliaia di persone che praticano sport in modo saltuario
16419 – 16017 = +402 migliaia di persone che praticano sport solo qualche volta

Le **variazioni assolute** del numero di coloro che fanno sport al 2020 rispetto al 2019 (e quindi nel 2020 rispetto al tempo precedente) sono:

15837 – 15605 = +232 migliaia di persone che praticano sport in modo continuativo
5559 – 4905 = +654 migliaia di persone che praticano sport in modo saltuario
16419 – 17234 = –815 migliaia di persone che praticano sport solo qualche volta

Per sapere se dal 2011 al 2020 è incrementato maggiormente il numero di coloro che fanno sport in modo continuativo o il numero di coloro che praticano sport solo qualche volta possiamo confrontare le variazioni assolute? La risposta è NO.

- Le variazioni assolute, sia rispetto ad un tempo prefissato, sia rispetto al tempo precedente, non consentono un raffronto tra fenomeni differenti, poiché sono espresse nella stessa unità di misura dei valori a cui si riferiscono e risentono dell'ordine di grandezza.

Non è corretto confrontare le variazioni assolute di coloro che fanno sport in modo continuativo con quelle di coloro che fanno sport solo qualche volta, poiché l'ordine di grandezza delle seconde è molto più elevato. Ad esempio, per le variazioni rispetto al tempo base, si parte da 12717 persone per le variazioni dei praticanti sport in modo continuativo e da 16017 persone per i praticanti sport solo qualche volta. A maggior ragione non sarebbe lecito il confronto delle variazioni assolute di variabili in diverse unità di misura (ad esempio, il numero di coloro che praticano sport con le ore dedicate alla lettura di libri)

Per confrontare le variazioni di serie storiche diferenti si utilizzano i numeri indice, che mostrano i valori percentuali, confrontabili, delle variazioni. I numeri indice si distinguono tra:

- Numeri indice a base fissa
- Numeri indice a base mobile

Numeri indice a base fissa

- Si pone uguale a 100 il valore assunto da ciascuna serie storica in un tempo detto base (solitamente il primo) e si riferiscono tutti i dati a tale valore.
- Si sostituisce alla scala originaria dei tempi una nuova scala, indicata con t, ponendo $t = 1$ per il primo tempo, $t = 2$ per il secondo, ..., $t = T$ per l'ultimo.
- I n.i. a base fissa sono definiti come il quoziente, moltiplicato per 100, tra il valore x_t assunto dalla variabile nel tempo t ed il valore al tempo 1, x_1:

$$\frac{x_t}{x_1} \cdot 100 \quad t = 1, ..., T$$

- Sottraendo 100 ad un n.i. a base fissa si ottiene la variazione percentuale della variabile rispetto al tempo base. Infatti la variazione percentuale al tempo t rispetto al tempo base 1, può ottenersi anche rapportando la rispettiva variazione assoluta al valore della serie al tempo 1:

$$\left(\frac{x_t - x_1}{x_1}\right) \cdot 100 = \frac{x_t}{x_1} 100 - 100$$

Numeri indice a base mobile

- Sono definiti come quozienti, moltiplicati per 100, tra termini consecutivi della serie storica:

$$\frac{x_t}{x_{t-1}} \cdot 100$$

- In una serie storica riferita a T tempi vi sono $(T-1)$ n.i. a base mobile $\Rightarrow$ il n.i. a base mobile per il primo tempo non può essere calcolato

La tabella riporta i numeri indice a base fissa (base 2011) e quelli a base mobile della serie storica del numero di coloro che praticano sport in modo continuativo

ANNI	tempi	Praticano sport in modo continuativo	Numeri indice base fissa	Numeri indice base mobile
2011	1	12717	100,0	-
2012	2	12743	100,2	100,2
2013	3	12602	99,1	98,9
2014	4	13582	106,8	107,8
2015	5	14013	110,2	103,2
2016	6	14792	116,3	105,6
2017	7	14607	114,9	98,7
2018	8	15107	118,8	103,4
2019	9	15605	122,7	103,3
2020	10	15837	124,5	101,5

Notiamo come, ad esempio, nel corso di questi 10 anni i praticanti siano aumentati del 24,5%. L'ultimo anno c'è stato un aumento dell'1,5% e c'è stata una flessione solo negli anni dal 2012 al 2013, con un decremento dell'-1,1% e dal 2016 al 2017 (-1,3%). L'incremento inquesto 10 anni di coloro che praticano sport continuativamente è stato maggiore dell'incremento di coloro che particano sport solo qualche volta (uguale a $402/16017 \times 100 = 2,51\%$).

Relazione fondamentale tra numeri indice a base mobile e a base fissa
I numeri indice a base mobile al tempo *t* sono il rapporto tra il numero indice a base fissa al tempo e il numero indice a base fissa al tempo $(t-1)$:

$$\frac{x_t}{x_{t-1}} \cdot 100 = \left(\frac{\frac{x_t}{x_1}100}{\frac{x_{t-1}}{x_1}100} \right) \cdot 100$$

Questa relazione permette di passare da una serie di numeri indice all'altra. Ad esempio, il numero indice a base mobile al 2019 (pari a 103,3) può essere calcolato dal rapporto 122,7/118,8 x100.

Tasso medio di variazione

Il calcolo del tasso medio di variazione risponde alla domanda "quale variazione media percentuale ha avuto il fenomeno ogni anno per arrivare dal valore rilevato al tempo *1* al valore registrato al tempo *T*, l'ultimo della rilevazione?
Il tasso medio è proprio quel tasso che, se applicato ad ogni tempo, permette di arrivare dal valore x_1 al valore x_T.
La formula di calcolo è:

$$\left[\sqrt[T-1]{\frac{x_T}{x_1}} - 1 \right] \cdot 100$$

Si calcola la radice $T-1$ del numero indice a base fissa al tempo T diviso 100, quindi si sottrae 1 e si moltiplica per 100.
Se la serie è annuuale si parla di tasso medio annuale, se mensile di tasso medio mensile, se semestrale di tasso medio semestrale, etc…
Calcoliamo i tassi medi di variazione (tassi medi annui) di tutte le serie storiche riportate nella tabella:

- $\left[\sqrt[9]{1,245} - 1 \right] \times 100 = +2,47 \rightarrow$ il numero di coloro che praticano sport in modo continuativo, dal 2011 al 2020 è aumentato in media del 2,47% all'anno

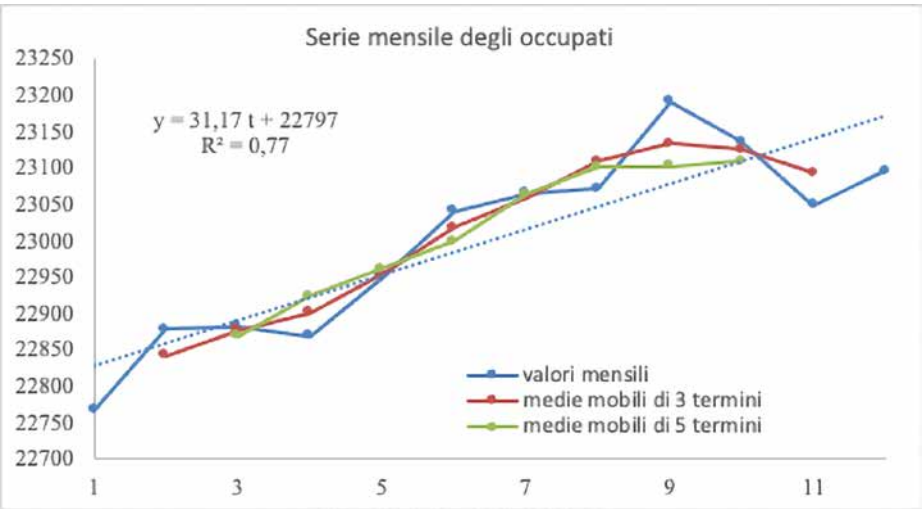
- $\left[\sqrt[9]{0,946}-1\right]\times 100=-0,62\rightarrow$ il numero di coloro che praticano sport in modo saltuario, dal 2011 al 2020 è diminuito in media dello 0,62% all'anno
- $\left[\sqrt[9]{1,025}-1\right]\times 100=+0,28\rightarrow$ il numero di coloro che praticano sport solo qualche volta, dal 2011 al 2020 è aumentato in media dello 0,28% all'anno
- $\left[\sqrt[9]{0,892}-1\right]\times 100=-1,26\rightarrow$ il numero di coloro che non praticano sport, dal 2011 al 2020 è diminuito in media dell'1,26% all'anno

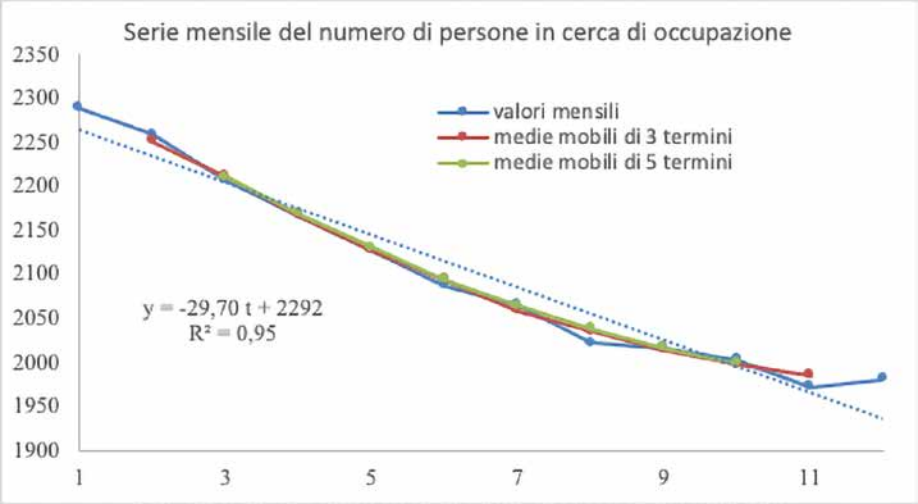
AL LAVORO!

LA TABELLA SEGUENTE (FONTE: DATI ISTAT) RIPORTA IL NUMERO DI OCCUPATI E DISOCCUPATI IN CERCA DI OCCUPAZIONE IN ITALIA DA OTTOBRE 2021 A SETTEMBRE 2022.
CHE COSA POSSIAMO DIRE SULLA SITUAZIONE OCCUPAZIONALE IN ITALIA NEL PERIODO CONSIDERATO?

Mese	t	Occupati (in migliaia)	In cerca di occupazione (in migliaia)
Ottobre 2021	1	22767	2289
Novembre 2021	2	22878	2259
Dicembre 2021	3	22883	2206
Gennaio 2022	4	22869	2167
Febbraio 2022	5	22950	2127
Marzo 2022	6	23040	2087
Aprile 2022	7	23065	2065
Maggio 2022	8	23071	2022
Giugno 2022	9	23191	2016
Luglio 2022	10	23136	2003
Agosto 2022	11	23049	1972
Settembre 2022	12	23095	1980

Suggerimento:





- Dall'analisi del trend si evince che nei 12 mesi rilevati c'è stato un andamento crescente del numero di occupati ed una decrecita del numero di disoccupati in cerca di occupazione, molto decisi prima del tempo 8 (maggio 2022) e meno marcati negli ultimi mesi.
- Si stima un incremento medio di circa 31 occupati aogni mese ed un decremento medio di circa 30 disoccupati ogni mese. Il trend del numero di dìisoccupati ha un andamento decisamente lineare: la bontà di adattamento della retta è prossima ad 1.
- Nell'arco dei 12 mesi considerati gli occupati sono aumentati dell'1,44% e le persone in cerca di occupazione sono diminuite del 13,5%. Si è avuto quindi una variazione maggiore nel numero di persone in cerca di occupazione
- Mediamente, ogni mese gli occupati sono aumentati dell' 0,13% mentre le persone in cerca di occupazione sono diminuite dell'1,3%

CAPITOLO 3. LA STATISTICA PER I DATI SOCIALI

In questo capitolo ci concentreremo sui dati sociali, una tipologia di dati impiegata per lo studio di fenomeni demografici, politici ed economici, e sul loro impiego nella statistica inferenziale.
La disciplina che si occupa dello studio di questa tipologia di dati è la scienza sociale, una branca della scienza dedicata allo studio delle società e delle relazioni tra gli individui. All'interno della scienza sociale troviamo discipline come la sociologia, l'antropologia, l'economia, la politica, la geografia umana e la comunicazione.
La tecnica di rilevazione più diffusa per i dati sociali è l'inchiesta campionaria (*survey*), che prevede la somministrazione di un questionario ad un campione di individui rappresentativo della popolazione. Le domande del questionario hanno la funzione di operativizzare determinate caratteristiche che il ricercatore vuole indagare. L'inchiesta campionaria consente di raccogliere una serie informazioni in modo uniforme, semplificandone la codifica e l'analisi. Spesso, infatti, le risposte a disposizione dei partecipanti sono limitate (es. risposta multipla) o facilmente codificabili come le scale di valutazione (es. scala *Likert*[2]). Al termine della fase di rilevazione, i dati vengono organizzati in una matrice di dati, che riporta nelle righe gli individui osservati e nelle colonne le variabili del questionario. All'interno di ogni cella della matrice troviamo quindi un *dato*, ovvero un valore corrispondente alla risposta data da un rispondente per una particolare variabile.
L'indagine campionaria può essere condotta sia dal ricercatore che da soggetti terzi. Nel primo caso si parla di dati primari, raccolti appositamente dal ricercatore per lo scopo dello studio, mentre nel secondo caso si parla di dati secondari. I dati secondari vengono raccolti soggetti terzi come amministrazione pubbliche o enti statistici (es. Istituto Nazionale di Statistica – ISTAT) durante la loro ordinaria attività amministrativa oppure con il fine di studiare un determinato fenomeno sociale, come nel caso dei censimenti della popolazione.
A livello nazionale, l'ISTAT si occupa della rilevazione di dati riguardanti diversi fenomeni sociali, economici e commerciali. Alcune delle aree di interesse dell'ente nazionale di statistica riguardano ambiente ed energia, popolazione e famiglie, sanità e salute, educazione, giustizia e criminalità, agricoltura, turismo, cultura, commercio e trasporti. Allo stesso modo, l'*Eurostat* è l'organo statistico della Commissione Europea e si occupa di raccogliere dati ed indicatori sugli stati membri dell'Unione Europea.
Inoltre, tra le inchieste campionarie più conosciute a livello internazionale troviamo l'*European Social Survey* (ESS), un'indagine biennale sugli atteggiamenti ed i modelli di comportamento dei cittadini europei, il *Trends in International Mathematics and Science Study* (TIMSS) ed il *Progress of International Literacy study* (PIRLS), che valutano le abilità degli studenti a livello internazionale in matematica, scienza e lettura, ed il *Programme for International Student Assessment* (PISA), il cui obiettivo è di valutare i sistemi educativi nel mondo attraverso la rilevazione delle abilità scolastiche di studenti di 15 anni.
Al giorno d'oggi la maggior parte dei dati secondari accessibili al pubblico sono reperibili online, sui portali dei diversi enti che si occupano della loro rilevazione, come per esempio: ISTAT, World Bank, Organization for Economic Co-operation and Development (OECD), ed Eurostat.

Nel corso del capitolo verranno introdotte le tecniche di statistica inferenziale che permettono di indagare le differenze tra gruppi di individui e prevedere il valore di una particolare variabile partendo dai dati a disposizione.

3.1 INFERENZA STATISTICA

Come descritto in precedenza, i dati sociali fanno riferimento a fenomeni demografici, economici, e politici. Tuttavia, lo studio di questi fenomeni sociali è spesso caratterizzato da popolazioni di interesse ad alta numerosità,

2 La scala Likert è una tecnica psicometrica che valuta l'opinione del rispondente riguardo un determinato fenomeno attraverso un set ordinato di risposte. Ad esempio, una scala Likert [1-5] è composta dai valori (1-2-3-4-5), dove 1 corrisponde a "Totalmente in disaccordo" e 5 corrisponde a "Totalmente d'accordo".

rendendo difficile per il ricercatore raggiungere tutti i soggetti.
Per questo motivo, i ricercatori si avvalgono di tecniche di ***statistica inferenziale***, che consentono di studiare aspetti specifici della popolazione attraverso un campione rappresentativo (es. individui, famiglie, imprese). Le tecniche di inferenza statistica permettono infatti di estendere le conclusioni derivanti dall'analisi dei dati campionari all'intera popolazione di riferimento, rendendo lo studio dei fenomeni sociali alla portata dei ricercatori. Una delle tecniche dell'inferenza statistica è il test di verifica d'ipotesi, che andremo ora ad approfondire.

3.2 VERIFICA D'IPOTESI

Il test di verifica d'ipotesi è una procedura inferenziale utilizzata per trarre conclusioni circa un dato parametro della popolazione (es. media μ) che risulta incognito sulla base di stime campionarie. A tale proposito, nell'ipotesi il parametro (μ) assume un valore ipotizzato (μ_0), e sulla base della stima del campione si potrà stabilire se la differenza tra i due sia dovuta a cause accidentali o sistematiche.
Questa procedura consente dunque di tradurre la domanda di ricerca in ipotesi statistiche. Ad esempio, supponiamo di essere interessati all'altezza media delle donne adulte in Italia (μ). Dai dati ISTAT, risulta che l'altezza media femminile nel nostro paese è pari a 165 cm. La nostra popolazione di riferimento è quindi costituita da tutte le donne che hanno compiuto 18 anni residenti in Italia. Per poter stimare il parametro della popolazione (μ), estraiamo un campione di 25 donne e calcoliamo la media campionaria ($\bar{x}$ = 162.5 cm). A questo punto possiamo utilizzare il test di ipotesi per verificare se la media campionaria differisce in modo statisticamente significativo dalla media della popolazione di riferimento. Per fare ciò andremo a seguire i quattro step per la verifica di ipotesi:

a) Formulazione delle ipotesi
b) Definizione di un criterio decisionale
c) Calcolo della statistica test
d) Decisione sulle ipotesi formulate

a) ***Formulazione delle ipotesi***
La prima ipotesi, che solitamente comprende la relazione di uguaglianza o nessun effetto, viene detta ***ipotesi nulla*** (H_0). L'ipotesi nulla è un'affermazione su un parametro della popolazione, come la media, che si presume sia vera. Per contro, l'***ipotesi alternativa*** (H_1) è un'affermazione che contraddice direttamente l'ipotesi nulla, affermando che il valore effettivo di un parametro della popolazione è inferiore, superiore o non uguale al valore indicato nell'ipotesi nulla.
A seconda della formulazione di H1, possiamo avere tre tipi di test:

$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	Test bilaterale (a due code)
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	Test unilaterale (a coda a destra)
$H_0 : \mu = \mu_0$	$H_1 : \mu > \mu_0$	Test unilaterale (a coda a sinistra)

Nonostante le informazioni siano rilevate sul campione, le ipotesi sono espresse con riferimento al corrispettivo parametro della popolazione perché si è interessati all'intera popolazione di riferimento.

b) ***Definizione di un criterio decisionale***
In questa fase andremo a definire un criterio che da utilizzare per prendere decisioni. Tale criterio prende il nome di ***livello di significatività*** (α), che determina se un dato risultato può essere considerato statisticamente significativo. Consiste nella probabilità di rifiutare l'ipotesi nulla, quando in realtà è vera. I livelli di significatività più comunemente utilizzati sono 1% e 5% (convenzionale).
Un concetto strettamente legato al livello di significatività è l'intervallo di confidenza, ovvero un intervallo costruito attorno ad una statistica (es. media) in modo tale che sia nota la probabilità che il parametro appartenga all'intervallo stesso. Tale probabilità è detta **livello di confidenza** ed indicato con $(1-\alpha)$% dove α è la probabilità che il parametro si trovi al di fuori dell'intervallo di confidenza.

c) ***Calcolo della statistica test***

La statistica test è un valore che viene utilizzato utilizzando i dati a disposizione sulla popolazione e sul campione. Il risultato della statistica test viene utilizzato per prendere decisioni con riferimento all'ipotesi nulla formulata al punto 1. In presenza di variabili ***continue***, a seconda delle informazioni a disposizione sulla popolazione di riferimento, possiamo distinguere le seguenti casistiche:

- grandi campioni (n>30) e varianza (σ^2) nota → Z-test
- piccoli campioni (n<30) e varianza (σ^2) ignota → T-test

Per poter svolgere questi test (1) le osservazioni devono essere indipendenti, (2) il campione deve essere estratto in modo casuale dalla popolazione, (3) la popolazione di riferimento deve avere una distribuzione normale.

Z-test

Nel primo caso possiamo utilizzare la statistica test con distribuzione normale standard (Z). Per calcolare la statistica 'z' utilizziamo la formula:

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ = media campionaria
μ_0 = media ipotizzata della popolazione
σ = deviazione standard della popolazione
n = dimensione campionaria
** il fattore ($\sigma/\sqrt{n}$) rappresenta l'errore standard

La soluzione della formula fornisce il numero di deviazioni standard, per cui la media di un campione è superiore o inferiore alla media della popolazione indicata nell'ipotesi nulla.

T-test

Nel caso di campioni piccoli e varianza ignota andremo ad utilizzare la statistica test con distribuzione *t di Student* con 'n-1' gradi di libertà (df)[3].

$$t_{df} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \qquad df = n - 1$$

$\bar{x}$ = media campionaria
μ_0 = media ipotizzata della popolazione
s = deviazione standard della popolazione
n = dimensione campionaria

d) ***Decisione sulle ipotesi formulate***

Una volta calcolata la statistica test, si può quindi confrontare il valore del test con i valori critici relativi al livello di significatività (α). Il valore critico (nelle tabelle) della statistica del test è il valore della statistica per il quale si rifiuta l'ipotesi nulla al livello di significatività dato.

| Statistica test | > valore critico → rifiutiamo l'ipotesi nulla (H_0)
| Statistica test | < valore critico → accettiamo l'ipotesi nulla (H_0)

3 *Gradi di libertà ('degrees of freedom')*: rappresentano il numero di informazioni indipendenti libere di variare nel calcolo di una determinata stima di una statistica.

Ipotesi alternativa	Area di rifiuto
$H_1: \mu > \mu_0$	$z_{\alpha,} + \infty$
$H_1: \mu < \mu_0$	$- \infty, z_{\alpha}$
$H_1: \mu \neq \mu_0$	$(- \infty, z_{\alpha/2})$ e $(z_{\alpha/2,} + \infty)$

Z-test:

Il valore $z_{\alpha/2}$ di Z che viene scelto per costruire l'intervallo di confidenza è chiamato valore critico. A ciascun livello di confidenza (1-α) corrisponde un diverso valore critico. I valori critici della distribuzione Z più comunemente utilizzati per i test a una e due code sono:

Livello di significatività (α)	Tipo di test	
	Unilaterale	*Bilaterale*
0.05	+1.645 or -1.645	± 1.96
0.01	+2.33 or -2.33	± 2.58
0.001	+3.09 or -3.09	± 3.30

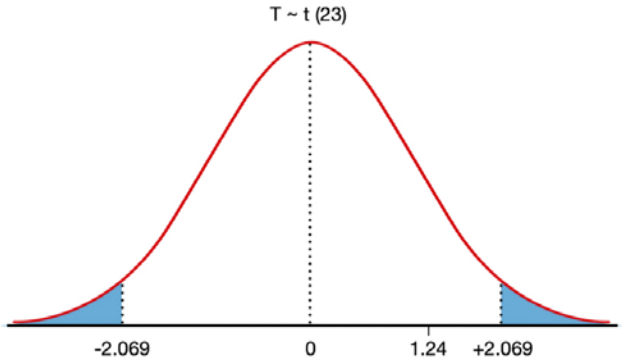
T-test:

Una volta ottenuto il valore della statistica del test T, cerchiamo il valore critico ($t_{\alpha,df}$) nelle tavole della distribuzione *t di Student*, in base al livello di significatività (α) e ai gradi di libertà (df).

Distribuzione t di Student											
prob. cum	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
una coda	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
due code	1	0.50	0.40	0.40	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
...	...	...	...	...	...	...	...	...	...	...	...
23	0.000	0.685	0.858	1.060	1.319	1.714	**2.069**	2.500	2.807	3.485	3.768
24	0.000	0.685	8.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745

Per individuare il valore critico sulle tavole statistiche occorre sapere i gradi di libertà (df = n-1), il livello di significatività (α), ed il tipo di test (unilaterale o bilaterale). Ad esempio, per un test bilaterale con df = 23 e α = 5%, il valore critico è pari a 2.069. Se confrontato con una statistica test t_{23} = 1.24, la nostra decisione sarà quella di accettare l'ipotesi nulla, in quanto il valore ricade nell'area di accettazione (Fig. 1).

Figura 1. Distribuzione t di Student



AL LAVORO!

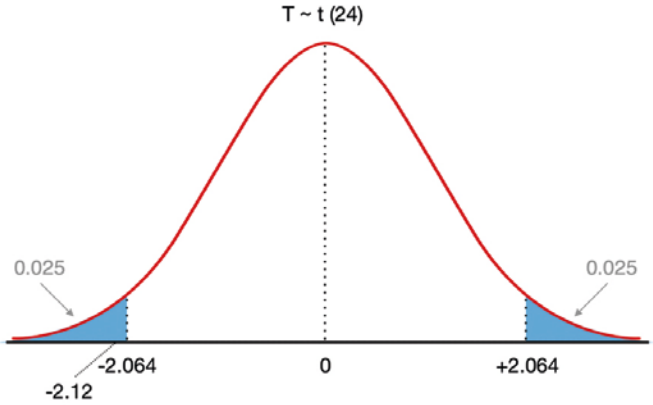
A. NEL NOSTRO ESEMPIO RELATIVO ALL'ALTEZZA MEDIA DELLE DONNE ITALIANE, OPTANDO PER UN TEST BILATERALE LE NOSTRE IPOTESI SARANNO:
H_0: $\mu = 168$ H_1: $\mu \neq 168$

B. SCEGLIAMO ORA UN LIVELLO DI SIGNIFICATIVITÀ PARI AL 5% ($\alpha = 0.05$).

C. CALCOLIAMO LA STATISTICA TEST. TRATTANDOSI DI UN CAMPIONE PICCOLO (N<30) E NON CONOSCENDO LA VARIANZA (σ^2), OPTEREMO PER UN T-TEST. RICORDIAMO CHE PER UN CAMPIONE DI 25 DONNE, $\bar{X} = 162.5$ CM, LA DEVIAZIONE STANDARD CAMPIONARIA È $S = 5.88$ CM, E IL VALORE IPOTIZZATO PER LA POPOLAZIONE $\mu_0 = 165$ CM.

$$t_{24} = \frac{162.5 - 165}{\frac{5.88}{\sqrt{25}}} = -2.12$$

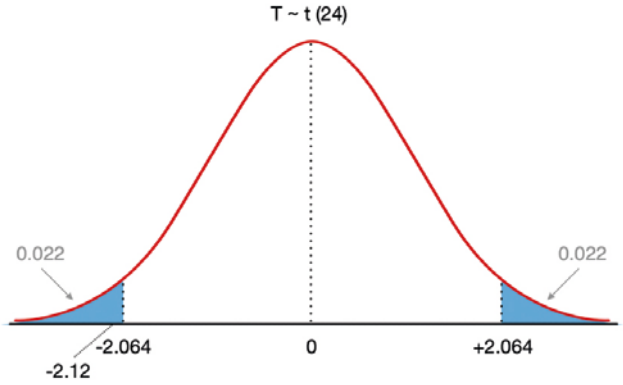
D. DALLE TAVOLE DELLA DISTRIBUZIONE T DI *STUDENT* TROVIAMO IL VALORE CRITICO PER UN TEST A DUE CODE CON $\alpha = 0.05$ E 'N-1' GRADI DI LIBERTÀ (DF=24). IL VALORE CRITICO È PARI A $T_{0.025,24} = 2.064$, PERTANTO RIFIUTIAMO L'IPOTESI NULLA DI NON-DIFFERENZA TRA LE MEDIE.
LA DIFFERENZA TRA LE DUE MEDIE È QUINDI STATISTICAMENTE SIGNIFICATIVA.



Approccio alternativo: il *p-value*

Il *p-value* rappresenta la probabilità di ottenere una particolare misura statistica (es. media campionaria) se l'ipotesi nulla fosse vera per la popolazione. Come ogni probabilità, il *p-value* è compreso tra 0 e 1 e non può mai essere negativo. Questo valore viene fornito durante la verifica di ipotesi dalla maggior parte dei software statistici ma può anche essere approssimato partendo dalle tavole delle distribuzioni. Una volta ottenuto il *p-value* lo possiamo utilizzare per prendere decisioni nella verifica di ipotesi.
In particolare, se il *p-value* è inferiore al livello di significatività (α), si può rifiutare l'ipotesi nulla (H_0) e concludere che la differenza è statisticamente significativa. Viceversa, se il *p-value* è maggiore di alpha, accettiamo l'ipotesi nulla. Tanto più basso è il *p-value,* quanto più forte è l'evidenza dei dati contro H_0.

Nell'**esempio** precedente relativo all'altezza media delle donne italiane, con una statistica t = -2.12 e df = 24 otteniamo un *p-value* = 0.044. Siccome il *p-value* è inferiore al livello di significatività $\alpha = 0.05$, rifiutiamo l'ipotesi nulla (H_0), giungendo quindi alla stessa conclusione. In questo caso, trattandosi di un test bilaterale, dividendo per 2 il *p-value* otteniamo 0.022, ovvero la probabilità associata a ciascuna delle due aree di rifiuto.



3.3 TEST A DUE CAMPIONI

Finora ci siamo concentrati su test di verifica d'ipotesi svolti su un singolo campione, ma questo procedimento può anche essere esteso per verificare le differenze tra due campioni. Nei test a campione singolo, la statistica campionaria viene confrontata con un valore standard della popolazione. Nei test su due campioni, le statistiche campionarie di due campioni vengono messe a confronto per determinare la differenza statistica tra di essi. Analogamente a quanto già visto per i test su campione singolo, anche in questo caso possiamo condurre un Z-test o un T-test, a seconda delle informazioni a disposizione.

Z-test a due campioni

Può essere applicato quando la varianza (o deviazione standard)[4] della popolazione è nota ed il campione è sufficientemente grande (n>30). In questo caso la statistica test è calcolata come:

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\bar{x}_1$ = media campione 1
μ_1 = media ipotizzata popolazione 1
σ^2_1 = varianza popolazione 1
n_1 = dimensione campione 1

$\bar{x}_2$ = media campione 2
μ_2 = media ipotizzata popolazione 2
σ^2_2 = varianza popolazione 2
n_2 = dimensione campione 2

T-test a due campioni

T-test: utilizzato nei casi in cui la varianza di entrambe le popolazioni è sconosciuta ed il campione di piccole dimensioni (n<30). Nel confronto tra medie di due campioni possiamo poi distinguere le seguenti tre casistiche:

- la varianza è sconosciuta e si presume che sia *uguale* per entrambe le popolazioni;
- la varianza è sconosciuta e si presume sia *diversa* per ciascuna popolazione;
- i due campioni derivano da due osservazioni distinte sulla stessa unità (test a campioni *accoppiati*).

Il T-test a due campioni con ***varianza uguale*** è calcolato come segue:

$$t_{df} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad df = n_1 + n_2 - 2$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

4 Ricordiamo che la deviazione standard (σ) non è altro che la radice quadrata della varianza (σ^2).

$\bar{x}_1$ = media campione 1
μ_1 = media ipotizzata popolazione 1
s^2_1 = varianza campione 1
s^2_p = varianza dei campioni accoppiati
n_1 = dimensione campione 1

$\bar{x}_2$ = media campione 2
μ_2 = media ipotizzata popolazione 2
s^2_2 = varianza campione 2
n_2 = dimensione campione 2

Il T-test a due campioni con ***varianza diversa***, anche detto *Welch's t-test*, si esegue con la formula:

$$t_{df} = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \qquad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Infine, la statistica test per il T-test a due campioni con ***accoppiati*** si ottiene con la formula:

$$t_{df} = \frac{\overline{x_d}}{\frac{s_d}{\sqrt{n}}} \qquad df = n - 1$$

$\bar{x}_d$ = media delle differenze tra campioni
s_d = deviazione standard delle differenze tra campioni
** il fattore $(sd/\sqrt{n})$ rappresenta l'*errore standard*

AL LAVORO!

ESERCIZIO 1: UN RICERCATORE È INTERESSATO AD INDAGARE LA DIFFERENZA NELLA MEDIA TRA I KM PERCORSI UTILIZZANDO CARBURANTE DIESEL E BLUDIESEL (A PARITÀ DI LITRI DI CARBURANTE). AL FINE DI VERIFICARE SE ESISTE UNA DIFFERENZA STATISTICAMENTE SIGNIFICATIVA, RILEVA SU 14 AUTO I KM PERCORSI CON ENTRAMBI I CARBURANTI. LE INFORMAZIONI ESTRATTE DAI CAMPIONI SONO RIPORTATE IN TABELLA:

	Carburante (n=14)	
	Diesel	*BluDiesel*
media	$\bar{x}_{diesel} = 217.77$	$\bar{x}_{bludiesel} = 232.31$
deviazione standard	$s_{diesel} = 13.67$	$s_{bludiesel} = 22.57$

- ***Di che tipo di test si tratta? Perché?***

Dal momento che le informazioni rilevate sui due carburanti derivano dalle stesse unità statistiche, si tratta di un T-test a campioni accoppiati.

- ***Formulare l'ipotesi nulla e alternativa (test bilaterale).***

$H_0 : \mu_d = 0$ $\quad$ $H_1 : \mu_d \neq 0$

(la differenza tra campioni è pari a 0)

- ***Calcolare la statistica test e confrontarla con il valore critico (α = 0.05).***

Differenza tra medie: $\overline{x}_d = 217.77 - 232.31 = -14.54$
Deviazione standard della differenza tra campioni: $s_d = 21.41$

Calcoliamo la statistica test con 'n-1' gradi di libertà:

$$t_{13} = \frac{-14.54}{\frac{21.41}{\sqrt{14}}} = -2.54 \qquad df = 13$$

Sulle tavole, con df = 13 e α = 0.05 troviamo il valore critico $t_{0.025,13} = 2.16$.
Siccome $|-2.54| > 2.16$, rifiutiamo l'ipotesi nulla di non-differenza, concludendo che esiste una differenza statisticamente significativa tra i km percorsi con i due diversi tipi di carburante.
La stessa conclusione può essere ottenuta confrontando il *p-value* (0.024) al livello di significatività. Anche in questo caso, 0.024 < 0.05, pertanto rifiutiamo H_0.

AL LAVORO!

ESERCIZIO 2: SUPPONIAMO DI VOLER VERIFICARE SE UOMINI E DONNE DORMONO LA STESSA QUANTITÀ DI ORE. A TALE PROPOSITO, RACCOGLIAMO I DATI SU DUE CAMPIONI: IL CAMPIONE DEGLI UOMINI (N=19) PRESENTA UNA MEDIA DI ORE DI SONNO DI 6.76, IL CAMPIONE DELLE DONNE (N=22) INVECE REGISTRA UNA MEDIA PARI A 6.70.

	Ore di sonno	
	Uomini	*Donne*
dimensione	$n = 19$	$n = 22$
media	$\bar{x}_{uomini} = 6.76$	$\bar{x}_{donne} = 6.70$
deviazione standard	$s_{uomini} = 1.06$	$s_{donne} = 1.34$
varianza	$s^2_{uomini} = 1.12$	$s^2_{donne} = 1.80$

- ***Di che tipo di test si tratta? Perché?***

Si tratta di due campioni distinti, per i quali possiamo assumere una varianza uguale.

- ***Formulare l'ipotesi nulla e alternativa (test bilaterale).***

$H_0 : \mu_1 = \mu_2$ (i due campioni sono uguali) $\qquad H_1 : \mu_1 \neq \mu_2$

- ***Calcolare la statistica test e confrontarla con il valore critico (α = 0.05).***

$$s_p^2 = \frac{(19-1)1.12 + (22-1)\,1.80}{19+22-2} = 1.49$$

$$t_{39} = \frac{6.76 - 6.70}{\sqrt{1.49\left(\frac{1}{19}+\frac{1}{22}\right)}} = 0.153 \qquad df = 39$$

Dalle tavole, 39 gradi di libertà e e α = 0.05 troviamo il valore critico $t_{0.025,39} = 2.02$.
La statistica test è inferiore al valore critico, pertanto ricade nell'area di accettazione di H_0 (le due medie sono uguali). Anche il *p-value* associato (0.88) suggerisce la stessa conclusione.

3.4 TEST SU PIÙ CAMPIONI: ANOVA

Nell'insieme di tecniche di statistica inferenziale troviamo l'ANOVA (*Analysis Of Variance*), che consente il confronto delle medie tra due o più gruppi sulla base della varianza.

Questa tecnica consente infatti di indagare le differenze tra campioni confrontando la variabilità interna ai gruppi con la variabilità tra gruppi. Questo processo determina se i gruppi fanno parte di una popolazione più ampia o di popolazioni separate con medie diverse.

ANOVA ad una via (one-way ANOVA)

Per eseguire l'ANOVA ad una via è necessario disporre di una variabile ***dipendente continua*** e di una variabile ***indipendente categorica***[5] che suddivide i dati in gruppi da confrontare. Inoltre, le osservazioni devono essere indipendenti, la varianza dei gruppi approssimativamente uguale e ciascun gruppo deve avere la stessa dimensione (n).
In gergo statistico, le variabili indipendenti categoriche prendono il nome di *fattori*. Le categorie dei fattori sono detti 'livelli', che serviranno per creare i diversi gruppi. Le medie dei livelli dei fattori sono le medie della variabile dipendente associata a ciascun livello del fattore.
Ad esempio, in un esperimento il fattore 'metodo di insegnamento' potrebbe avere i seguenti tre livelli: 'metodo base', 'metodo 1', e 'metodo 2'. Il test ANOVA determinerà se le medie della variabile dipendente (punteggio test) per ciascun livello sono diverse.

	Variabile indipendente		
	Metodo base ***(n = 12)***	***Metodo 1*** ***(n = 12)***	***Metodo 2*** ***(n = 12)***
Punteggio test (var. dipendente)	$\bar{x}_1 = 8.5$	$\bar{x}_2 = 7.9$	$\bar{x}_3 = 8.3$

Possiamo quindi formulare le ipotesi da verificare:
H_0: $\mu_1 = \mu_2 = \mu_3$ (le medie dei gruppi sono tutte uguali)
H_1: almeno una media è diversa

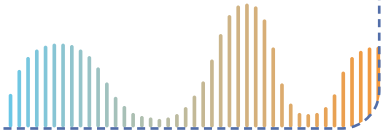
Successivamente, il test ANOVA utilizzerà l'F-test per determinare se la variabilità tra le medie dei gruppi è maggiore della variabilità delle osservazioni all'interno dei gruppi. Se questo rapporto è sufficientemente grande, si può concludere che non tutte le medie sono uguali.

$$F = \frac{\textit{varianza tra gruppi}}{\textit{varianza interna ai gruppi}}$$

La varianza tra gruppi ('*between*') consiste nella differenza tra la media della variabile dipendente e la media di ciascun gruppo. La varianza interna ai gruppi, detta anche varianza '*within*', è data dalla differenza tra la media e le osservazioni di ciascun gruppo. Avendo a disposizione più gruppi, questi valori fanno riferimento alla media delle somme dei quadrati.
Infine, il valore di F viene confrontato con il valore critico trovato sulle tavole della distribuzione di Fisher con gradi di libertà e livello di significatività corrispondenti.
Un risultato significativo nel test ANOVA indica che almeno una media differisce, ma non specifica quale. Per identificare quali differenze tra coppie di medie sono statisticamente significative, è necessario eseguire un'analisi *post-hoc*, che non approfondiremo in questo capitolo.

AL LAVORO!

UN DIETOLOGO È INTERESSATO A TESTARE LE DIFFERENZE NELLA PERDITA DI PESO MEDIA DEI SUOI PAZIENTI, IN BASE ALLE DIETE SEGUITE DA CIASCUNO DI ESSI. UTILIZZA QUINDI LA TECNICA ANOVA PER TESTARE SE C'È UNA DIFFERENZA STATISTICAMENTE SIGNIFICATIVA TRA I KG PERSI IN MEDIA (VAR. DIPENDENTE) PER OGNI TIPO DI DIETA (VAR. INDIPENDENTE). PER CIASCUN TIPO DI DIETA RACCOGLIE DATI SU 30 PAZIENTI, OTTENENDO I SEGUENTI RISULTATI:



5 Il termine 'categorica' indica variabili non-metriche, come quelle nominali e ordinali.

		Perdita di peso	
Tipo di dieta	**n**	**Media**	**Varianza**
Dieta A	30	6.41	0.96
Dieta B	30	6.72	1.23
Dieta C	30	7.13	0.97

Di seguito sono riportati i risultati del test ANOVA.

	df	Varianza (media dei quadrati)	F	*p-value*
Tra gruppi (*between*)	2	3.94	3.73	0.028
In gruppi (*within*)	87	1.06		
Totale	**89**			

Le ipotesi da testare saranno: H_0: $\mu_1 = \mu_2 = \mu_3$ e H_1: almeno una media è diversa
La statistica test è uguale a:

$$F = \frac{3.94}{1.06} = 3.73$$

Impostando un livello di significatività pari al 5% ($\alpha = 0.05$), dalle tavole della distribuzione F troviamo un valore critico ≈ 3.1, che suggerisce di rifiutare l'ipotesi nulla.
Anche il *p-value*, che risulta inferiore al livello di significatività ($0.028 < 0.05$), conferma l'esistenza di una differenza tra la perdita di peso media tra i vari tipi di dieta.

ANOVA a due vie (two-way ANOVA)

Il test ANOVA a due vie testa le differenze tra le medie di una variabile dipendente, i cui dati sono raggruppati sulla base di due variabili indipendenti categoriali (fattori).
Nei risultati dell'ANOVA a due vie, oltre agli effetti principali dei fattori, sono riportati anche gli effetti di interazione tra di loro. Il termine di interazione indica se l'effetto di una delle variabili indipendenti sulla variabile dipendente è lo stesso per tutti i valori dell'altra variabile indipendente (e viceversa). In presenza di due fattori, formuleremo un'ipotesi nulla per ciascuno di essi:

H_{01}: tutti i livelli del fattore 1 hanno la stessa media.

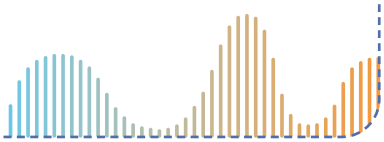
H_{02}: tutti i livelli del fattore 2 hanno la stessa media.

H_1: almeno una media è diversa

Inoltre, sarà necessario definire anche le ipotesi per il termine di interazione:
H_{0i}: non c'è interazione tra fattori H_{1i}: c'è interazione tra fattori

AL LAVORO!

UN RICERCATORE VUOLE SCOPRIRE SE ESISTONO DELLE DIFFERENZE TRA IL REDDITO ANNUO (VAR. DIPENDENTE) IN BASE AL GENERE (VAR. INDIPENDENTE 1) E ALLA LAUREA CONSEGUITA (VAR. INDIPENDENTE 2). DI SEGUITO SONO RIPORTATE LE STATISTICHE DESCRITTIVE DEI DATI RACCOLTI.



UOMINI		Reddito annuo*	
Laurea	**n**	**Media**	**Varianza**
Economia	20	77	33
Ingegneria	20	62.5	24.6
Medicina	20	69	23.9

**Dati espressi in migliaia di euro.*

DONNE		Reddito annuo*	
Laurea	**n**	**Media**	**Varianza**
Economia	20	74.1	16.9
Ingegneria	20	54.8	29.3
Medicina	20	67.2	30.3

**Dati espressi in migliaia di euro.*

Formuliamo ora le nostre ipotesi nulle e alternative:
H_{0G}: il genere non produce differenze di reddito
H_{0L}: l'indirizzo di studio non genera differenze di reddito
H_1: almeno una media è diversa

H_{0i}: non c'è interazione tra fattori H_{1i}: c'è interazione tra fattori

Il test ANOVA a due vie produce i seguenti risultati:

	df	Varianza (media dei quadrati)	F	*p-value*	F critico
Genere	1	515.67	19.58	0.000	3.92
Laurea	2	2854.59	108.37	0.000	3.08
Interazione	2	96.25	3.65	0.029	3.08
Totale	**119**				

I valori critici di 'Genere' e 'Laurea' risultano nettamente inferiori alla statistica test, pertanto rifiutiamo entrambe le ipotesi nulle H_{0G} e H_{0L}. Esiste quindi una differenza statisticamente significativa tra i redditi medi a seconda del genere e della laurea conseguita.
Inoltre, anche il termine di interazione tra fattori risulta significativo (*p-value* < 0.05). L'effetto principale del genere sul reddito medio cambia da un indirizzo di studio all'altro. Nello specifico, gli uomini hanno un reddito medio più alto, e questo effetto dipende dall'indirizzo di studio.

3.5 REGRESSIONE LINEARE

La regressione lineare ***semplice*** viene utilizzata per modellare la relazione bivariata tra due variabili, una indipendente (x) ed una dipendente (y), stimando l'intensità dell'effetto che la prima esercita sulla seconda. La variabile dipendente deve essere numerica, mentre quella indipendente può essere di qualsiasi tipo, sia categorica che numerica (contrariamente all'ANOVA).
Attraverso il modello di regressione, possiamo quindi prevedere il valore di una variabile di risposta (y), sulla base al valore di una variabile di ingresso (x). Inoltre, il modello consente anche di stimare *nuovi* valori della variabile dipendente per specifici valori della variabile indipendente.

Per ogni osservazione '*i*' con ($i = 1, \ldots, n$), l'equazione di regressione lineare viene definita come:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i \text{ (formulazione alternativa)}$$

y = variabile dipendente
α = intercetta (costante)
β = coefficiente di regressione
x = variabile indipendente (predittore, variabile esplicativa)
ε = termine di errore

Nello specifico, l'intercetta (α) fornisce il valore medio di '*y*' quando il predittore è uguale a zero.
Il coefficiente (β) indica la pendenza della retta di regressione, ovvero quanto cambia '*y*' per una variazione unitaria di '*x*'. Il segno del coefficiente indica la direzione della relazione tra le due variabili.
Il termine d'errore (ε) rappresenta la variazione della variabile dipendente che la variabile indipendente non spiega.

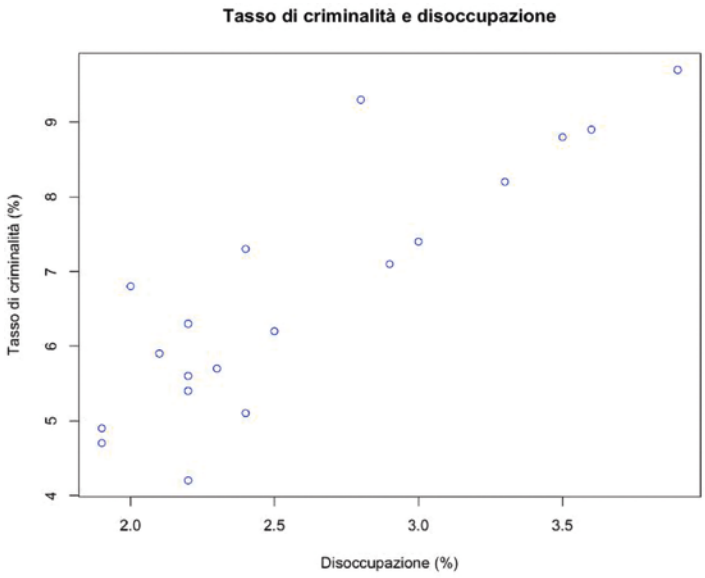
- La regressione lineare non è simmetrica in termini di '*x*' ed '*y*'. Ovvero, scambiando '*x*' ed '*y*' si otterrà un modello di regressione diverso ($x = y$) rispetto all'originale.

Requisiti per il modello di regressione lineare.

- ***Linearità***: la relazione tra '*x*' e la media di '*y*' è lineare;
- ***Indipendenza***: le osservazioni sono indipendenti l'una dall'altra;
- ***Normalità***: i residui del modello hanno una distribuzione normale.
- ***Omoscedasticità***: la varianza dei residui è la stessa per qualsiasi valore di '*x*';
- Le variabili indipendenti non sono correlate con il termine di errore (ε).

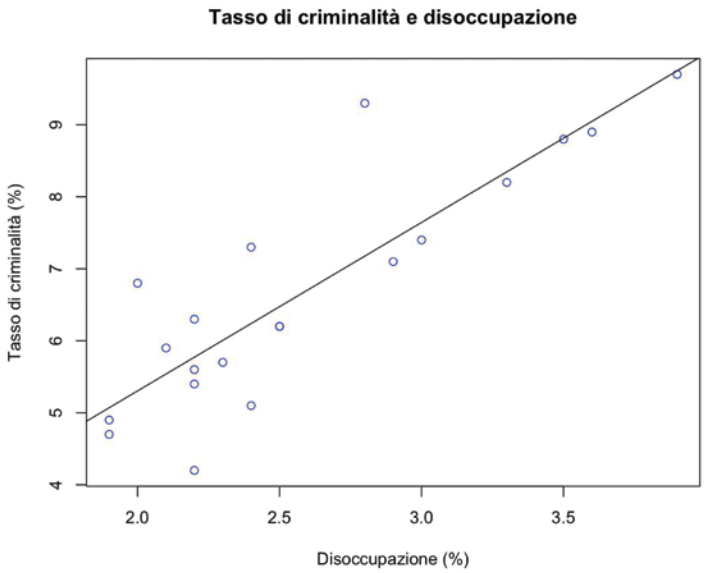
Esempio: supponiamo di voler investigare la relazione tra tasso di criminalità (y) e tasso di disoccupazione (x) in 20 città Europee. La relazione tra le due variabili può essere visualizzata in Figura 2.

Figura 2. Grafico a dispersione



Il grafico a dispersione (*scatter plot*) indica se esiste una relazione tra le variabili '*y*' e '*x*'.
Seguendo la prassi statistica, l'asse *Y* mostra la variabile dipendente, il tasso di criminalità (%). L'asse *X* (orizzontale) mostra la variabile indipendente, ovvero il tasso di disoccupazione (%).
Nel nostro esempio, possiamo notare una relazione positiva tra variabili abbastanza pronunciata, che può essere approssimata con una retta di regressione (Figura 3).

Figura 3. Retta di regressione lineare



Il modello di regressione traccia una linea attraverso le osservazioni che minimizza la loro distanza complessiva dalla linea stessa. Più precisamente, il metodo dei minimi quadrati (*Ordinary Least Squares – OLS*) minimizza la somma delle differenze al quadrato tra le osservazioni e la retta di regressione. L'equazione di regressione lineare derivante dai dati è la seguente:

$$criminalità_i = \alpha + \beta\ (disoccupazione)_i + \epsilon_i$$

La stima dei parametri tramite metodo dei minimi quadrati produce i risultati in tabella.

	Coefficienti	Errore standard	Valore t	*p-value*
Intercetta	0.623	0.8658	0.720	0.481
Tasso di disoccupazione (%)	2.340	0.3262	7.174	0.000

L'equazione di regressione che ne deriva è:

$$criminalità_i = 0.623 + 2.34\ (disoccupazione)_i + \epsilon_i$$

Dai risultati emerge che il tasso di criminalità medio in assenza di disoccupazione è pari a 0.623%. Il coefficiente positivo (β) indica che la relazione bivariata è positiva ed il *p-value* suggerisce che tale relazione è statisticamente significativa (*p-value*< α = 0.05).
Come possiamo notare, per valutare la significatività dei parametri stimati viene utilizzato il T-test introdotto nella Sezione 2. In questo caso, la statistica t viene calcolata con la formula:

$$t_{df} = \frac{parametro}{errore\ standard\ del\ parametro}$$

Anche in questo caso, è necessaria la formulazione delle ipotesi. Ad esempio, le ipotesi relative all'intercetta saranno:
$H_0 : \beta_0 = 0$
$H_1 : \beta_0 \neq 0$

Infine, per valutare la bontà di adattamento della retta ai valori di 'y' e 'x', utilizziamo il coefficiente di determinazione (R^2). In questo caso, $R^2 = 0.726$, ovvero il 72.6% della varianza del tasso di criminalità può essere spiegata dal tasso di disoccupazione.

REGRESSIONE MULTIPLA

Contrariamente alla regressione lineare semplice, che prevede una sola variabile esplicativa, la regressione multipla consente l'inclusione di più variabili indipendenti.

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

Oltre alle assunzioni classiche del modello semplice, le variabili indipendenti devono essere indipendenti l'una dall'altra, altrimenti si incorre nel problema di multicollinearità.
Ogni variabile esplicativa dovrebbe avere un contributo indipendente alla variabile dipendente.
In questo caso, per valutare la significatività statistica dei parametri stimati utilizziamo il test F (Sezione 4). Questo test viene eseguito congiuntamente su tutti i coefficienti del modello di regressione.

Regressione lineare vs ANOVA
Confrontiamo ora la regressione lineare ed il modello ANOVA. Nella regressione lineare semplice, sia la variabile di risposta che il predittore sono continui. Nell'ANOVA, invece, la risposta è continua, ma la variabile esplicativa (fattore) è nominale.
La regressione ci fornisce un modello statistico che ci permette di prevedere una risposta a diversi valori della variabile dipendente, compresi i valori non inclusi nei dati originali. L'ANOVA misura lo spostamento medio della risposta per le diverse categorie del fattore. Di conseguenza, viene generalmente utilizzata per confrontare le medie dei diversi livelli del fattore.

FOCUS: ERRORI, RESIDUI E R^2

L'**errore** (o disturbo) di un valore osservato è la deviazione del valore osservato dal valore vero (non osservabile) di una quantità di interesse (per esempio, una media della popolazione), mentre il **residuo** di un valore osservato è la differenza tra il valore osservato e il valore stimato della quantità di interesse (ad esempio, una media campionaria). I residui rappresentano quindi le stime campionarie dell'errore per ogni osservazione.

Nella **regressione lineare i** residui sono positivi se si trovano al di sopra della retta di regressione e negativi se si trovano al di sotto della retta di regressione.
Il termine residuo (errore) per ogni osservazione è:

$$\varepsilon_i = y_i - \widehat{y}_i$$

La somma dei residui è sempre uguale a zero, assumendo che la retta di regressione sia effettivamente la retta ottimale.

Coefficiente di determinazione (R^2)
Il coefficiente di determinazione (R^2), misura quanto un modello di regressione lineare si adatti a un insieme di dati. In altre parole, rappresenta la percentuale della varianza della variabile dipendente (y) che può essere spiegata dalla variabile indipendente (x).
Al fine di calcolare il coefficiente di determinazione utilizzeremo tre tipi diversi di somma dei quadrati: somma dei quadrati di regressione (*sum of squares due to regression - SSR*), somma dei quadrati dei residui (*sum of squares error - SSE*) e somma dei quadrati totali (*sum of squares total - SST*).

Somma dei quadrati di regressione (SSR)
Somma delle differenze tra il valore stimato dal modello e la media della variabile dipendente.
Rappresenta la variazione spiegata attribuibile alla relazione tra 'x' e 'y'.

$$SSR = \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2$$

$\bar{y}$ = media della variabile dipendente
$\hat{y}$ = valore stimato dal modello per l'osservazione 'i' (i = 1,… , n)

Somma dei quadrati dei residui (SSE)
L'errore è la differenza tra il valore osservato e il valore stimato dal modello.
Fornisce la variazione attribuibile a fattori diversi dalla relazione tra 'x' e 'y'.

$$SSE = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = \sum_{i=1}^{n} \left(\varepsilon_i\right)^2$$

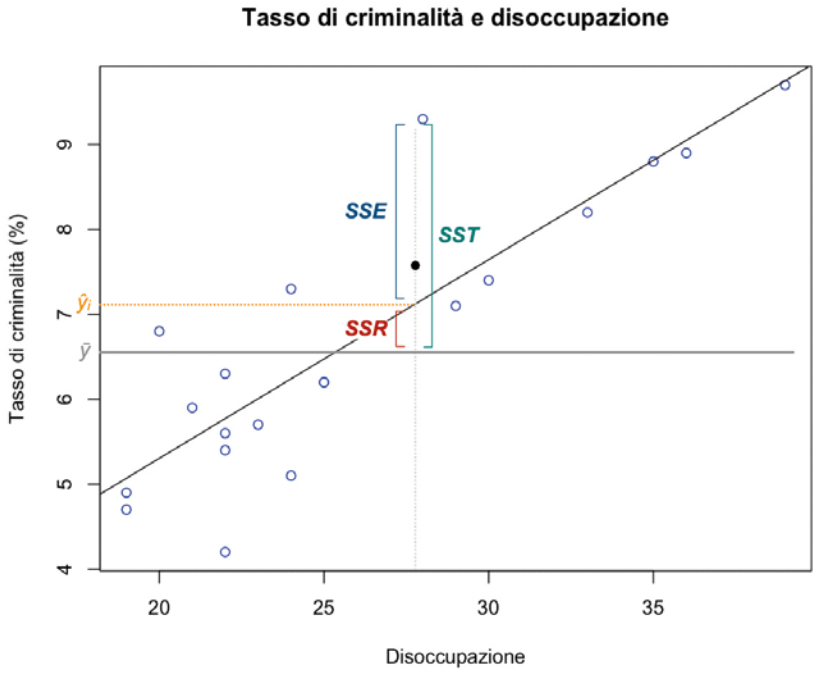
Somma dei quadrati totali (SST)
Somma dei quadrati delle differenze tra la variabile dipendente e la sua media. Misura la variazione dei valori y intorno alla loro media.

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SST = SSR + SSE$$

Nella Figura 4 sono riportati graficamente le tre somme dei quadrati.

Figura 4. Somme dei quadrati



Possiamo quindi calcolare il coefficiente di determinazione (R^2) come:

$$\boldsymbol{R^2} = \frac{SSR}{SST}$$

Oppure, utilizzando la formula alternativa:

$$\boldsymbol{R^2} = 1 - \left(\frac{SSE}{SST}\right)$$

Il valore di R^2 può variare da 0 a 1. Un valore di 0 indica che la variabile di risposta non può essere spiegata dalla variabile predittiva. Un valore di 1 indica che la variabile di risposta può essere perfettamente spiegata senza errori dalla variabile predittiva.
Nota bene: nella regressione lineare semplice, il coefficiente di determinazione corrisponde al coefficiente di correlazione di Pearson (r) al quadrato.

Adjusted R^2
Il coefficiente di determinazione R^2 aumenta sempre con l'inclusione di variabili indipendenti. Pertanto, nel caso di regressioni lineari multiple, occorre calcolare una versione del coefficiente che corregge l'inclusione di predittori aggiuntivi.

$$\overline{\boldsymbol{R^2}} = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$$

L'R^2 aggiustato tiene conto del numero di variabili esplicative (k) rispetto al numero di osservazioni (n).

CAPITOLO 4. LA STATISTICA PER I DATI SPAZIALI ED AMBIENTALI

I dati di natura **spaziale** sono dati di cui conosciamo la loro **locazione nello spazio** e si incontrano in diversi ambiti; noi ci concentriamo su due tipi particolari: i dati **ambientali** e i dati **epidemiologici**. Vedremo la struttura spaziale di entrambi e impareremo a trattarli dal punto di vista statistico.
Oggigiorno vengono raccolti un'infinità di dati per valutare lo stato dell'ambiente; ad esempio, se un fiume è a rischio esondazione è utile raccogliere dati sulla portata del fiume, per valutare il livello di inquinamento vengono raccolti dati sul particolato fine (PM10) che si trova nell'aria. La grande disponibilità di dati ha fatto si che l'interesse nelle metodologie statistiche per fenomeni ambientali sia cresciuto negli ultimi anni. Nel 1989, durante un meeting al Cairo, nasce *l'Environmetrics Society (TIES)*, che ha come obiettivo incoraggiare lo sviluppo e l'uso della statistica e di altri metodi quantitativi nelle scienze ambientali, nell'ingegneria ambientale, e nel monitoraggio e la protezione dell'ambiente. Da allora sono stati sviluppati tantissimi contributi interdisciplinari che propongono nuove metodologie statistiche per affrontare analisi di problemi ambientali. In Italia è attivo il *Gruppo di Ricerca per le Applicazioni della Statistica ai Problemi Ambientali (GRASPA)* con l'obiettivo di stimolare l'interazione tra ricercatori in campi diversi, favorendo così lo sviluppo di nuovi metodi statistici con applicazione a dati riguardanti l'ambiente.

La raccolta dei dati ambientali avviene in modi diversi, ma è comune sistemare delle stazioni di misurazione che raccolgono dati con una cadenza regolare, ad esempio ogni ora, ogni giorno, ecc. Questo da origine a una **matrice dei dati** dove le osservazioni non sono indipendenti tra di loro per via della struttura temporale nei dati, parliamo di **serie storiche**, che abbiamo già visto nel Capitolo 2. Spesso i dati vengono raccolti anche in diversi punti di una città, o su tutti i comuni di una certa regione, o per ogni paese dell'Europa. In Italia, l'Agenzia Regionale per la Prevenzione e l'Ambiente (ARPA) raccoglie dati tramite una rete di monitoraggio formata da centraline o stazioni distribuite su tutto il territorio italiano. La locazione di ogni centralina è data dalle sue coordinate geografiche (longitudine e latitudine). Questi sono dati con struttura spaziale, ed anche in questo caso abbiamo a che vedere con osservazioni dipendenti tra di loro!

D'altra parte, vengono raccolti dati sulla salute, ad esempio il numero di malati, ricoverati o morti a causa di una certa malattia nelle diverse regioni italiane. Questi dati sono oggetto di studio della **epidemiologia**, che si occupa della distribuzione e della frequenza delle malattie, ed anche in questo caso la statistica è di grande utilità! In Italia, l'ISTAT e il Servizio Sanitario Nazionale si occupano di raccogliere questi dati.

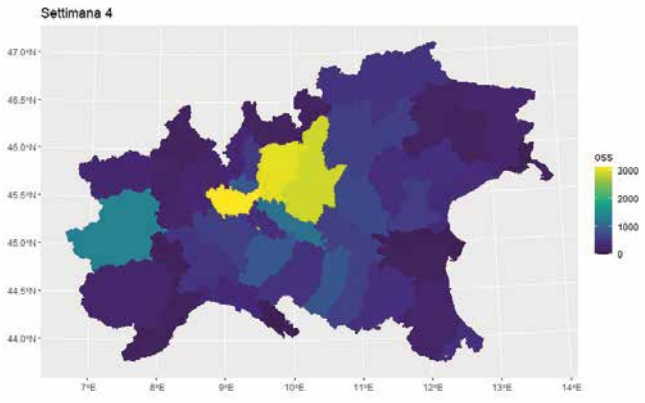
Sia nei dati ambientali che in quelli epidemiologici, ci aspettiamo correlazione spaziale, ma perché?
La prima legge della geografia, enunciata dallo geografo Waldo Tobler, dice che ``***le osservazioni ottenute in siti vicini tendono ad essere più simili tra loro rispetto alle osservazioni ottenute in siti lontani***''. Quindi, quanto più le osservazioni sono geograficamente vicine, tanto più tali misurazioni si rassomigliano. Se ignoriamo la presenza di **correlazione spaziale** abbiamo stime inaffidabili, ad es. dei coefficienti di regressione in un modello di regressione multipla mostrato nel Capitolo 3.
La **statistica spaziale** si occupa di studiare questo tipo di dati, che possiamo rappresentare su una mappa. Il concetto di autocorrelazione spaziale è generale e riguarda sia i fenomeni misurati su supporto spaziale discreto, chiamati **dati areali**, che sul continuo, detti **dati geostatistici**. Ora ne diamo descrizione precisa.

4.1 DATI AREALI

Consideriamo una regione di studio *A* ripartita in *n* aree (es. comuni). I dati areali sono dati raccolti su un supporto spaziale **discreto**, ad esempio il numero di malati di Covid in ogni provincia d'Italia. In genere i dati sulla salute sono **conteggi** di malati/decessi in forma aggregata a livello di comune, provincia, regione, ecc (archivi regionali, nazionali). I dati sono disponibili in forma aggregata sia a livello temporale (giorno, settimana, anno) che spaziale per ragioni di privacy, dato che questi dati sono considerati sensibili. Prendiamo l'esempio dei nuovi casi Covid ogni settimana, dove la settimana 1 corrisponde al 24 febbraio 2020. La nostra regione di studio *A* è il nord d'Italia, ed il numero di aree *n*= 47 (numero di province). La matrice dei dati ha la seguente forma:

Provincia	Settimana	Nuovi casi Covid	Popolazione
Piacenza	1	138	286433
Parma	1	35	454873
Reggio Emilia	1	4	529609
Modena	1	22	707119
…	…	…	…
…	…	…	…
Piacenza	2	341	286433
Parma	2	194	454873
Reggio Emilia	2	44	529609
Modena	2	60	707119
…	…	…	…

In questo esempio, ogni riga corrisponde ad una provincia del nord d'Italia. Come si può vedere nella tabella, abbiamo dati disponibili per diverse settimane (e quindi i dati di ogni provincia si potrebbero trattare come una serie storica!), ma noi ci concentriamo su una unica settimana qui (la settimana 4 come vedremo nella tabella sotto). La prima cosa utile da fare è visualizzare i dati su un grafico. Nel caso di dati areali, possiamo fare una **mappa coropletica**, ovvero una mappa dove ogni area *i=1,…,n* è colorata a seconda del valore della variabile di interesse. Noi siamo interessati ai "nuovi casi Covid", che rappresentano i conteggi **osservati** in ogni provincia.



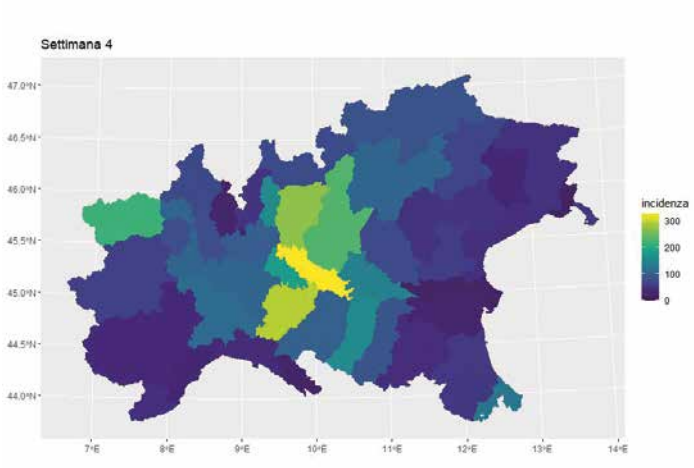
Se facciamo la mappa dei conteggi osservati però, non consideriamo affatto che il numero di persone a rischio (la popolazione) è diverso in ogni provincia, e la mappa sarà dominata dai conteggi nelle aree molto popolate dove il numero di persone a rischio è maggiore. Infatti, nella mappa sopra l'area con il maggiore numero di casi Covid corrisponde alla provincia di Milano, che però è anche quella con il maggiore numero di abitanti. Quindi è più utile mappare l'**incidenza** che considera il numero di nuovi casi in relazione alla dimensione demografica in ogni provincia:

incidenza = nuovi casi Covid / Popolazione a rischio

Possiamo vedere l'incidenza come il rischio assoluto (o la probabilità) di prendere il Covid in ogni provincia. Quindi aggiungiamo una nuova colonna nella nostra matrice dei dati:

Provincia	Settimana	Nuovi casi Covid	Popolazione	Incidenza
Piacenza	4	840	286433	0.002933
Parma	4	444	454873	0.000976
Reggio Emilia	4	824	529609	0.001556
Modena	4	600	707119	0.000849
…	…	…	…	…

Siccome i rischi assoluti sono numeri molto piccoli (per fortuna!), possiamo, per comodità, moltiplicarli per 100.000, cosi da mappare l'incidenza per 100.000 abitanti:



Cosa noti di diverso rispetto alla mappa precedente?

Nella mappa sopra possiamo vedere che l'incidenza più altra (colore giallo) si trova nella provincia di Cremona, quindi il rischio è superiore lì e non nella provincia di Milano! Vediamo anche che c'è correlazione o dipendenza spaziale, ovvero, provincie vicine tra di loro tendono ad essere colorate in modo simile.

Come valutare in questo caso il grado di dipendenza o correlazione spaziale? Esistono indicatori di autocorrelazione spaziale per dati areali; tali indicatori misurano il grado con cui osservazioni simili tendono a verificarsi in aree vicine/confinanti. In epidemiologia, questi indicatori possono aiutare ad identificare **cluster di malattia**, ovvero degli insiemi di aree vicine fra loro caratterizzate da rischio di malattia elevato o incrementato rispetto alla media. La presenza di cluster di malattia suggerisce che la malattia è contagiosa oppure che la malattia è associata a fattori di rischio ambientali che sono distribuiti in maniera non omogenea nella regione di studio. Uno di questi indicatori è **l'indice *I* di Moran,** che permette di investigare il grado di correlazione spaziale a livello globale, cioè su tutta la regione di studio *A*.
Date le osservazioni di un fenomeno *Y* in *n* aree, la similarità fra y_i e y_j è misurata come prodotto delle differenze delle osservazioni nelle aree *i* e *j* dalla media generale $\overline{y} = \sum_{i=1}^{n} y_i / n$

$$sim_{ij} = (y_i - \overline{y})(y_j - \overline{y})$$

L'**indice di Moran** è la media pesata delle similarità divisa per la varianza $s^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2 / n$

$$I = \frac{1}{s^2} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$$

dove w_{ij} è il peso da assegnare ad ogni similarità ed descrive la prossimità fra le aree *i* e *j*. Di solito, si omette la similarità di y_i con se stesso, assegnando di fatto peso nullo a w_{ii}. I pesi si raccolgono nella chiamata **matrice di vicinato**, una matrice di dimensione *n x n* dove ogni elemento w_{ij}:

$$w_{ij} = \begin{cases} 1 & \textit{se le aree i e j sono confinanti} \\ 0 & \textit{altrimenti;} \end{cases}$$

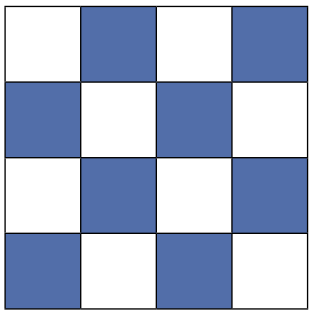
ogni riga/colonna della matrice di vicinato corrisponde ad una provincia. Vediamo, ad esempio, nella tabella sotto la matrice di vicinato per l'Emilia Romagna:

	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forli-Cesena	Rimini
Piacenza	0	1	0	0	0	0	0	0	0
Parma	1	0	1	0	0	0	0	0	0
Reggio Emilia	0	1	0	1	0	0	0	0	0
Modena	0	0	1	0	1	1	0	0	0
Bologna	0	0	0	1	0	1	1	0	0
Ferrara	0	0	0	1	1	0	1	0	0
Ravenna	0	0	0	0	1	1	0	1	0
Forli-Cesena	0	0	0	0	0	0	1	0	1
Rimini	0	0	0	0	0	0	0	1	0

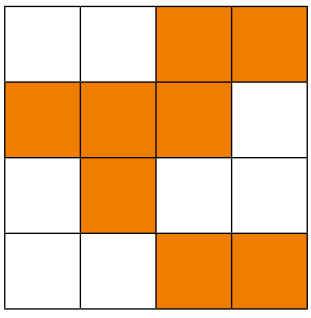
Come interpretare l'indice *I* di Moran? L'indice può prendere valori nell'intervallo [−1,1], dove:

- *I* < 0 indica dispersione, ovvero *Y* tende ad assumere valori molto diversi nelle aree vicine
- *I* = 0 indica assenza di correlazione spaziale, i valori alti/bassi sono distribuiti a caso spazialmente
- *I* > 0 indica correlazione spaziale positiva, ovvero *Y* tende ad assumere valori molto simili nelle aree vicine

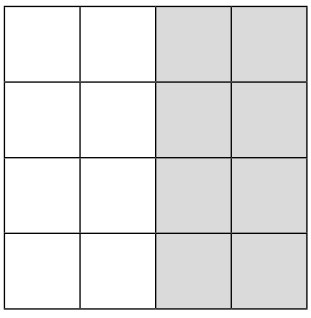
Moran *I* < 0



Moran *I* = 0



Moran *I* > 0



Calcoliamo l'indice *I* di Moran per i nostri dati Covid. Per prima cosa ci serve la media generale $\overline{y} = 0.000919$, poi dobbiamo calcolare le differenze $(y_i - \overline{y})$ ed infine i prodotti per poi pesarli con le informazioni della matrice di vicinato. Vediamo un piccolo esempio con 4 provincie:

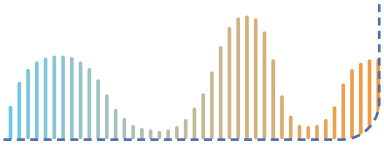
Provincia	Incidenza (y_i)	$(y_i - \overline{y})$
Piacenza	0.002933	0.002014
Parma	0.000976	0.000057
Reggio Emilia	0.001556	0.000637
Modena	0.000849	-0.000070

$w_{ij} * sim_{ij}$	Piacenza	Parma	Reggio Emilia	Modena
Piacenza	0	$1.15 \cdot 10^{-7}$	0	0
Parma	$1.15 \cdot 10^{-7}$	0	$3.63 \cdot 10^{-8}$	0
Reggio Emilia	0	$3.63 \cdot 10^{-8}$	0	$-4.46 \cdot 10^{-8}$
Modena	0	0	$-4.46 \cdot 10^{-8}$	0

Nella tabella a sinistra, abbiamo calcolato le differenze, mentre che la tabella a destra mostra, in arancione, i prodotti dei pesi e le similarità non nulli che vanno a contribuire nel numeratore del indice *I* di Moran. Per i dati Covid riferiti al nord Italia che stiamo usando come esempio, $I = 0.43$, per cui possiamo dire che c'è una correlazione spaziale positiva ma sostanzialmente debole nella incidenza del Covid nel nord Italia.

AL LAVORO!

LA TABELLA SOTTO MOSTRA I DATI DEI NUOVI CASI COVID NELLA SETTIMANA 5 DALL'INIZIO DELLA PANDEMIA NELLE 10 PROVINCIE DELLA TOSCANA. CALCOLA L'INDICE *I* DI MORAN PER L'INCIDENZA E COMMENTA IL RESULTATO.



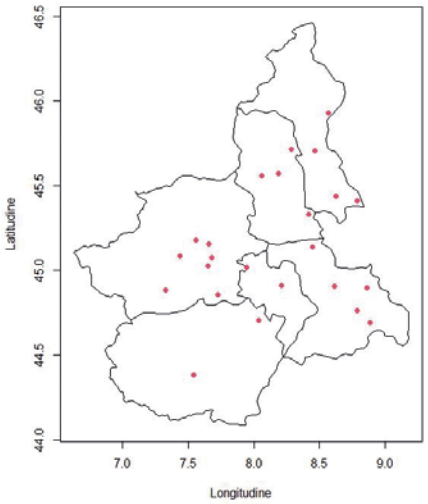
Provincia	Nuovi Casi Covid	Popolazione
Massa Carrara	214	191685
Lucca	311	382543
Pistoia	117	291697
Firenze	404	995517
Livorno	100	331877
Pisa	187	418122
Arezzo	136	339172
Siena	102	265179
Grosseto	128	219690
Prato	106	257073

- Per arrivare a calcolare l'indice *I* di Moran ti sarà utile:
  - Calcolare l'incidenza
  - Costruire la matrice di vicinato
  - Calcolare la media
  - Calcolare le differenze dalla media per ogni provincia
  - Calcolare i prodotti delle differenze per le coppie con peso diverso da zero

4.2 GEOSTATISTICA

La **geostatistica** è una branca della statistica spaziale con l'obiettivo di analizzare fenomeni che si manifestano su un supporto spaziale **continuo** $D \in \mathbb{R}^2$, ma che per motivi di opportunità vengono *osservati* in un numero finito di locazioni geografiche p_i, $i = 1, \ldots n$. In questo caso quindi abbiamo a disposizioni dati puntuali, cioè riferibili ad un punto dello spazio.

La geostatistica ha origine nell'ambito delle scienze minerarie, con l'obiettivo di fare **previsione** della distribuzione di minerale nel sottosuolo a partire da campionamenti del terreno; questa previsione spaziale in geostatistica viene chiamata **Kriging** e prende il nome da Krige, l'ingegnere minerario che lo sviluppò per primo.



Vediamone un esempio. Siamo interessati a monitorare i livelli di PM10 (µg/m^3) nella regione del Piemonte (quindi la regione del Piemonte, rappresentata nella mappa a sinistra, è il supporto spaziale continuo D in questo caso). Potenzialmente, il PM10 si potrebbe misurare su qualsiasi punto della mappa, ma quello che abbiamo a disposizione è il PM10 misurato solamente nei 24 punti segnati in rosso.

Come fare previsione del PM10 su tutto il Piemonte a partire dai dati che abbiamo a disposizione?

I **dati geostatistici** hanno due caratteristiche:

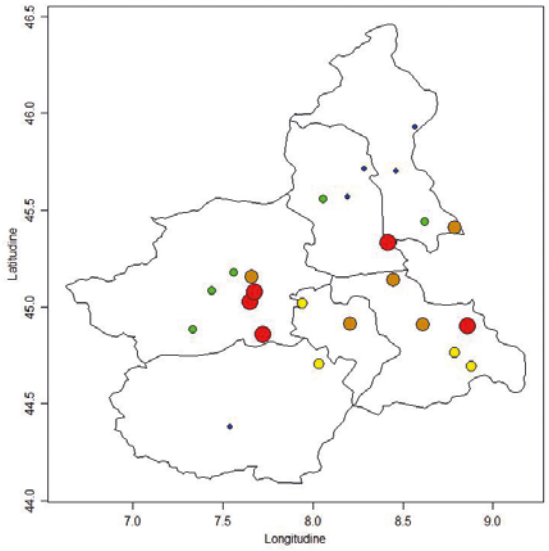
- per un dato sito i è di solito disponibile una sola osservazione
- le osservazioni del fenomeno in due siti distinti non sono indipendenti fra loro, bensì vi è dipendenza spaziale.

La matrice dei dati in questo caso ha la seguente forma:

Stazione	Longitudine	Latitudine	PM10
1	8.61	44.91	95,0
2	8.03	44.71	51,7
3	8.88	44.69	58,0
4	8.21	45.91	99,7
…	…	…	…

Nella mappa sopra abbiamo segnato in rosso la locazione spaziale dalle 24 stazioni di misurazione. Possiamo fare una mappa simile per mostrare, ad ogni stazione, il livello di PM10 misurato. Ci sono diverse opzioni, possiamo usare simboli diversi per indicare i diversi valori, lo stesso simbolo ma di grandezza diversa (dove la grandezza è proporzionale al valore del PM10), oppure diversi colori. Siccome tutti i valori osservati di PM10 sono diversi tra di loro, per fare il grafico conviene raggrupparli.

In questo caso, usiamo i quintili (Q_i, i=1,2,3,4) per definire 5 gruppi di valori, che riportiamo con colori e dimensioni diverse nella seguente mappa:



- gruppo 1 (blu): valori di PM10 < Q_1
- gruppo 2 (verde): valori di PM10 tra Q_1 e Q_2
- gruppo 3 (giallo): valori di PM10 tra Q_2 e Q_3
- gruppo 4 (arancione): valori di PM10 tra Q_3 e Q_4
- gruppo 5 (rosso): valori di PM10 > Q_4

Cosa osservi su questa mappa? I colori uguali tendono ad essere vicini tra di loro oppure sono disposti a caso?

Prima di fare previsione sui punti dove non ci sono osservazioni, dobbiamo valutare la correlazione spaziale. Indichiamo la correlazione spaziale tra due punti a distanza u come $\rho(u)$. Lo strumento per descrivere la correlazione spaziale in questo caso si chiama **variogramma**.
Il **variogramma** V associato alla coppia di punti $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$ è definito come:

$$V(p_1, p_2) = \frac{1}{2} Var\left(S(p_1) - S(p_2)\right)$$

dove $S(p_1)$ rappresenta la variabile PM10 nel punto p_1. Il variogramma $V(p_1, p_2)$ misura la forza della relazione lineare fra le variabili $S(p_1)$ e $S(p_2)$. Sotto certe assunzioni, il variogramma dipende dalla sola distanza euclidea $u = \| p_1 - p_2 \|$ tra i due punti, per cui possiamo vederlo come una funzione di u ed scrivere il variogramma $V(u)$ in:

$$V(u) = \sigma^2\left(1 - \rho(u)\right)$$

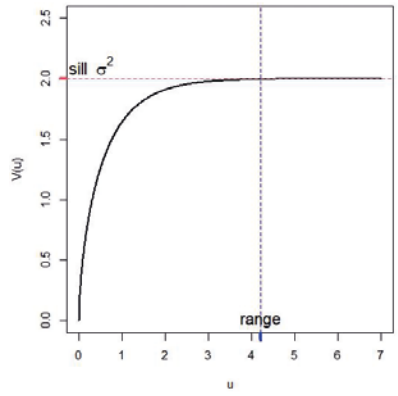
dove $\rho(u)$ è la cosiddetta funzione di correlazione spaziale. Ricordate che ci aspettiamo che ***osservazioni ottenute in siti vicini siano più simili tra loro rispetto a osservazioni ottenute in siti lontani;*** questo si traduce in una funzione di correlazione spaziale $\rho(u)$ che:

- tenderà ad 1 per punti molto vicini tra di loro (cioè, quando u, intesa come distanza tra due punti, tende a 0)
- tenderà a zero per punti lontani tra di loro (ovvero, man mano che u diventa grande)

In corrispondenza, il variogramma $V(u)$:

- tenderà a zero quando la distanza tra due punti u tende a 0
- tenderà a σ^2 (chiamato "*sill*") quando la distanza u tende ad infinito

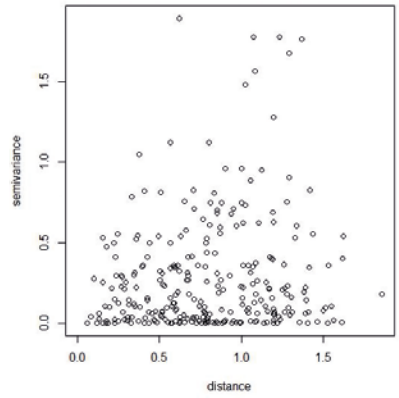
Il variogramma teorico ha la seguente rappresentazione grafica:



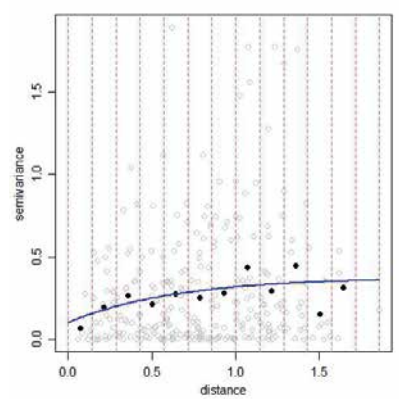
Come si può vedere nel grafico, da certo punto in poi (dal "range" in poi) il variogramma si appiattisce. Questo punto rappresenta la distanza a partire della quale la correlazione spaziale diventa nulla. La curva $V(u)$ che vedete nel grafico a sinistra ha una espressione analitica che dipende di alcuni parametri (sill e range).

A questo punto, con i dati osservati che abbiamo a disposizione, possiamo ricavare una **stima** del variogramma teorico che chiamiamo **variogramma empirico**. Il variogramma empirico è un grafico i cui valori di ordinata sono calcolati sulla base delle osservazioni $y_1, \ldots, y_n$:

- in ascissa: le distanze $u_{ij} = \| p_i - p_j \|$
- in ordinata: il valore $v_{ij} = \frac{1}{2}(y_i - y_j)^2$



Sul grafico a sinistra, troviamo il variogramma empirico per i nostri dati di PM10. A meno di dati osservati su una griglia regolare, le distanze fra coppie di punti saranno tutte diverse, per cui questo stimatore non è molto utile (la forma della nuvola di punti non assomiglia per niente alla forma del variogramma teorico che abbiamo visto prima). Per migliorare la leggibilità del variogramma empirico possiamo fare una media dei valori v_{ij}, per le coppie di siti la cui distanza rientra in un certo intervallo, come illustrato nel seguente grafico:



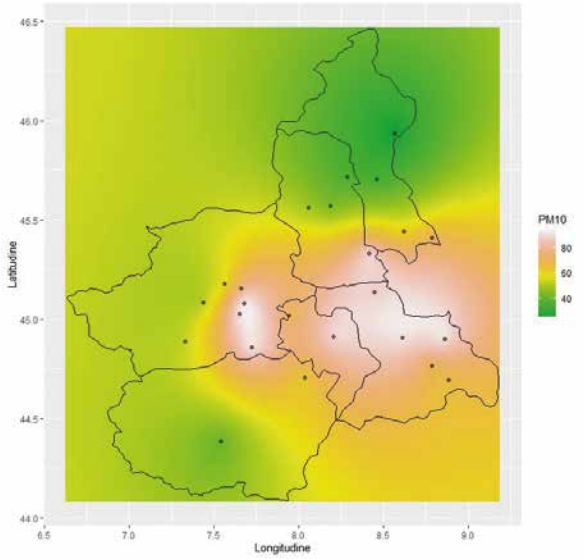
- dividiamo l'asse orizzontale in intervalli per definire le strisce segnate in rosso sul grafico
- calcoliamo la media dei valori v_{ij} che si trovano all'interno di ogni striscia, rappresentata dai punti neri

Il variogramma empirico "mediato" viene usato per stimare i parametri di un variogramma teorico e di conseguenza per stimare la funzione di correlazione spaziale $\rho(u)$. Ci sono diversi metodi di stima dei parametri, tra i cui il **"metodo dei minimi quadrati"** (ricordate che minimizza la somma dei residui al quadrato). Il variogramma teorico stimato è rappresentato dalla linea blu.

Ora siamo pronti per **prevedere** il PM10 (**kriging**) in un nuovo punto dove il PM10 non è stato osservato. Ma come possiamo usare al meglio i dati che abbiamo per fare previsione? La previsione y_0 viene calcolata come una somma pesata dei dati osservati y_i, $i = 1, \ldots, n$:

$$y_0 = \sum_{i=1}^{n} w_i y_i$$

I pesi w_i sono calcolati in modo da assegnare un maggiore peso a quei dati che sono più vicini al punto, cioè tenendo conto della correlazione spaziale $\rho(u)$ che abbiamo stimato tramite il variogramma.

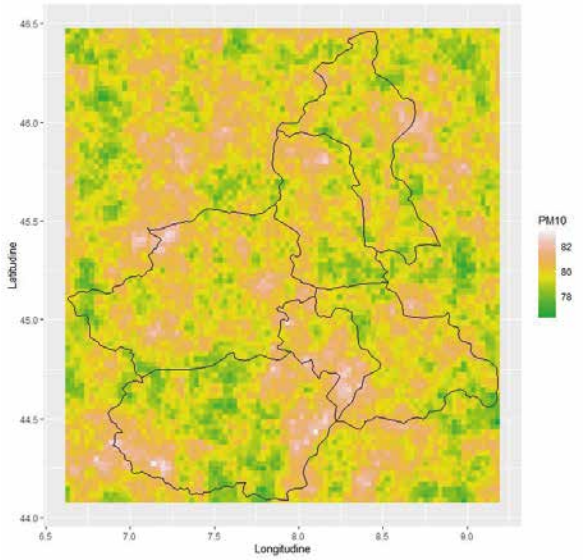


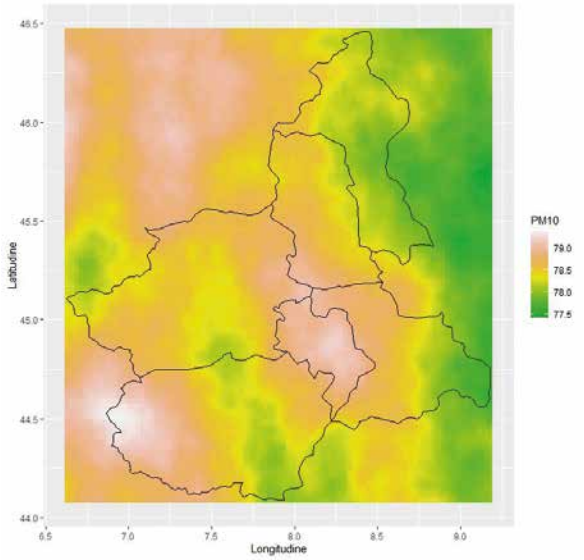
Possiamo pensare di fare previsione su una griglia molto fine che va a coprire tutto lo spazio di interesse (nel nostro caso, la regione del Piemonte), così da ottenere una mappa di PM10 su tutto il territorio come quella che vedete a sinistra.

- Ricorda che i valori di PM10 mostrati dalla mappa sono delle previsioni (gli abbiamo "calcolato" a partire dai dati osservati e la correlazione spaziale stimata) e non dei valori misurati, per cui saranno soggetti ad incertezza.

AL LAVORO!

SOTTO TROVI DUE MAPPE DI PM10 PER DUE GIORNI DIVERSI. SAPRESTI DIRE IN CHE GIORNO CI ASPETTIAMO UNA CORRELAZIONE SPAZIALE CON UN RANGE MAGGIORE? RICORDA CHE IL *RANGE* È LA DISTANZA A PARTIRE DELLA QUALE LA CORRELAZIONE SPAZIALE DIVENTA NULLA.





CAPITOLO 5. LA STATISTICA PER I BIG DATA

5.1 UN MONDO DI DATI

Proprio nell'istante in cui state leggendo questo testo, nel mondo viene generata un'enorme quantità di dati. La creazione di questa grande mole di dati è dovuta alla digitalizzazione della maggior parte dei processi, all'esistenza di diverse piattaforme di social network, ai numerosi dispositivi digitali e alla diffusione nell'uso del web. Ogni giorno anche voi navigate sul web, controllate la vostra posta elettronica, scrivete messaggi in molte chat e cliccate su diversi adds. Ogni mossa che eseguite online equivale alla creazione di dati.
Vi siete mai chiesti quanti dati vengono creati ogni giorno?
Si stima che il volume di dati creati, copiati e consumati a livello globale nel 2020 sia stato di 44 zettabyte, cresciuto a 79 zettabyte nel 2021 e si prevede che tale numero raddoppierà nel 2025 (fonte: Statista).

▸ Nota che se il megabyte è il multiplo di 10^6 del byte, lo zettabyte corrisponde al multiplo di 10^{21}!!

Ovviamente questo volume di dati è stato amplificato a causa della pandemia da coronavirus. Durante la pandemia sono stati chiusi uffici, scuole, ristoranti e altri stabilimenti e di conseguenza tante persone hanno trascorso più tempo su Internet per lavoro e tempo libero. Oggi il tasso di generazione dei dati è talmente elevato che supera la capacità delle tecniche di archiviazione dei dati esistenti.

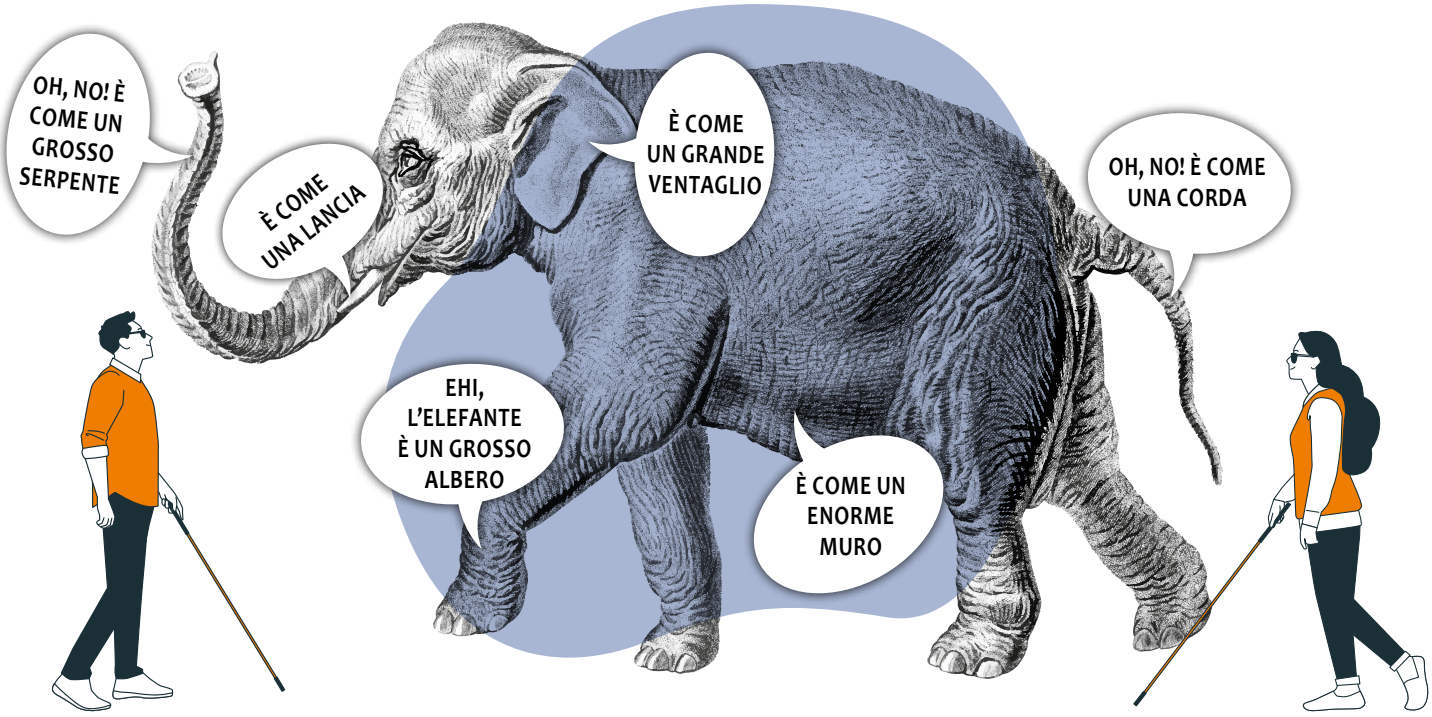
Quanti dati sono stati creati giornalmente nel 2020?

- Una ricerca condotta nel 2016 ha stimato che 1,7 MB è la quantità di dati creati ogni secondo per persona (Northeastern University)
- Ogni giorno sono stati creati 2,5 quintilioni di byte di dati (SG Analytics, 2020)
- Tutto ciò equivale a 10 milioni di dischi blu-ray, che una volta impilati sarebbero alti quanto due torri Eiffel messe insieme (Dihuni, 2020)
- I dati archiviati crescono 5 volte più velocemente dell'economia mondiale (Dihuni, 2020)
- Nel mese di agosto 2020, in un minuto sono stati scambiati 41.666.667 messaggi da utenti WhatsApp (Domo, 2020)
- Ogni minuto sono stati 404.444 gli utenti collegati in streaming su Netflix (Domo, 2020)
- amazon ha spedito 6.659 pacchi al minuto, evidenziando una crescita esplosiva dell'e-commerce nel 2020 (Domo, 2020)
- Gli utenti di posta elettronica hanno inviato 306,4 miliardi di e-mail al giorno nel 2020 (TechJury, 2020)
- Sono stati inviati 500 milioni di tweet al giorno, 5.787 tweet al secondo (TechJury, 2020)
- Su Google sono state effettuate 3,5 miliardi di ricerche (e-Learning Infographics, 2020)
- Su YouTube sono state caricate 300 ore di video al minuto (e-Learning Infographics, 2020)

5.2 COSA SONO I BIG DATA?

Per fornire un'intuizione di cosa sono i big data, vi proponiamo un racconto di origini indiane, intitolato "I ciechi e l'Elefante" che fu tradotto come poema in lingua inglese nel XIX secolo dallo scrittore John Godfrey Saxe.
Si racconta che sei ciechi abitavano in un villaggio. Un giorno gli abitanti del villaggio dissero loro: "oggi c'è un elefante nel villaggio". I ciechi non avevano idea di cosa fosse un elefante ma dissero: "anche se non saremmo in grado di vederlo, andiamo comunque a sentirlo". Così tutti andarono e toccarono l'elefante.

- "Ehi, l'elefante è un grosso albero", disse il primo cieco che gli toccò la gamba.
- "Oh, no! è come una corda", disse il secondo cieco che toccò la coda.
- "Oh, no! è come un grosso serpente", disse il terzo cieco che toccò la proboscide dell'elefante.
- "È come un grande ventaglio" disse il quarto cieco che toccò l'orecchio dell'elefante.
- "È come un enorme muro", disse il quinto cieco che toccò il ventre dell'elefante.
- "È come una lancia", disse il sesto cieco che toccò la zanna dell'elefante.



Cominciarono a discutere sulle caratteristiche dell'elefante e ognuno di loro insisteva per avere ragione. Un uomo saggio passando di lì, li vide e chiese loro: "Qual è il problema?". I ciechi risposero: "Non riusciamo ad essere d'accordo su com'è l'elefante". Ognuno di loro raccontò cosa pensava dell'elefante. Allora il saggio spiegò loro: "Tutti avete ragione. Il motivo per cui ognuno di voi lo definisce in modo diverso deriva dal fatto che ognuno di voi ha toccato una parte diversa dell'elefante. In realtà l'elefante ha tutte quelle caratteristiche che ciascuno di voi ha identificato".
Come l'elefante, anche i big data possiedono molteplici caratteristiche e possono essere soggetti a varie definizioni, letture ed interpretazioni a seconda della prospettiva nella quale sono osservati ed analizzati.
L'etimologia del termine "Big Data" risale alla metà degli anni Novanta quando fu utilizzato per la prima volta da John Mashey, *Chief Scientist* in pensione presso *Silicon Graphics*, per fare riferimento alla gestione e all'analisi di enormi insiemi di dati[6]. Successivamente Laney[7] ha caratterizzato i big data mediante tre caratteristiche, anche note come 3 V:

- **Volume**: sono big perché sono costituiti da enormi quantità di dati!
- **Velocity (velocità)**: sono dati creati in real-time!

6 Diebold F (2012) A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline. Available at: http://www.ssc.upenn.edu/fdiebold/papers/paper112/Diebold_Big_Data.pdf
7 Laney D (2001) 3D data management: Controlling data volume, velocity and variety. In: Meta Group. Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

- **Variety (varietà)**: sono dati strutturati, semi-strutturati e non strutturati, quindi essi sono molto variegati!

Da allora, altre caratteristiche sono state attribute ai Big Data, tra cui le seguenti:

- **Exhaustivity (esaustività)**: nella raccolta dati viene considerato un intero sistema o un'intera popolazione (n=tutti!) piuttosto che un singolo campione (*Mayer-Schonberger e Cukier, 2013*)
- **Fine-grained (a grana fine):** per quanto riguarda la risoluzione (*Dodge e Kitchin, 2005*)
- **Relationality (relazionalità)**: perché contengono campi comuni che abilitano la congiunzione di diversi set di dati (*Boyd e Crawford, 2012*)
- **Extensionality (estensionalità)**: perché si possono aggiungere o modificare facilmente nuovi campi (*Marz e Warren, 2012*)
- **Scaleability (unità di scala):** perché si possono aumentare rapidamente le dimensioni (*Marz e Warren, 2012*)
- **Veracity (veridicità)**: i dati possono essere disordinati, sporchi e contenere incertezza ed errori (*Marr, 2014*)
- **Value (valore)**: molte intuizioni possono essere testate sui dati; ne consegue che essi sono fonte di potenziale valore (*Marr, 2014*)
- **Variability (variabilità)**: sono dati il cui significato può essere costantemente mutevole in base al contesto (*McNulty, 2014*).

Uprichard[8] utilizza ulteriori parole, sempre con iniziale V, per descrivere i Big Data, tra le quali: "*versatility, volatility, virtuosity, vitality, visionary, vigour, viability, vibrancy... virility... valueless, vampire-like, venomous, vulgar, violating and very violent*".
Al contrario, Lupton[9] ha recentemente suggerito di abbandonare le suddette V-words per adottare 13 P-words nella sua proposta di descrizione dei Big Data: "*portentous, perverse, personal, productive, partial, practices, predictive, political, provocative, privacy, polyvalent, polymorphous and playful*".
Comunque vengano definiti, il messaggio che si vuole trasmettere è il seguente: i Big Data sono sostanzialmente diversi dai dati tradizionali per molteplici caratteristiche, che non si riducono alla sola dimensione o all'ampiezza. Alcune di esse sono sintetizzate nella seguente tabella.

	Small data	Big data
Volume	Da limitato a grande	Molto grande
Velocità	Fermi o lenti nel mutare	Rapida
Varietà	Da limitata ad ampia	Ampia
Esaustività	Campioni	Intere popolazioni
Risoluzione	Debole	A grana fine
Relazionalità	Da debole ad alta	Alta
Scala ed estensione	Medio-bassa	Alta

5.3 LA CLASSIFICAZIONE DEI BIG DATA

In un mondo di dati, la tipologia delle informazioni esistenti è molteplice. Comunque si vuole fornire una classificazione in 3 tipologie: dati provenienti dai social network, dati derivanti dai sistemi di business tradizionali e dati generati da macchine o sistemi. Qui di seguito si espone una spiegazione dettagliata.

1. **Social Network (informazioni da fonti umane).** Queste informazioni derivano dalla registrazione di esperienze umane, che in passato venivano raccolte in libri e opere d'arte, mentre oggi si usano fotografie,

8 Uprichard E (2013) Big data, little questions. Discover Society, 1 October. Available at: http://discoversociety.org/2013/10/01/focus-big-data-little-questions/

9 Lupton D (2015) The thirteen Ps of big data. The Sociological Life, 13 May. Available at: https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/

audio e video. Al presente le informazioni derivanti da fonti umane sono quasi interamente digitalizzate e vengono archiviate in diversi modi, dai personal computer ai social network. In questa tipologia i dati sono strutturati in modo approssimativo e spesso essi non sono completamente controllati. Alcuni esempi di tali dati sono i seguenti:

- Social Networks: Facebook, Twitter, Tumblr ecc.
- Blog e forum
- Immagini: Instagram, Flickr, Picasa ecc.
- Video: Youtube, ecc.
- Internet searches sui diversi motori di ricerca
- Dati da cellulari come messaggi di testo
- Mappe generate dagli utenti
- E-Mail

2. **Sistemi di business tradizionali.** Questi processi hanno lo scopo di registrare e monitorare gli eventi aziendali di interesse, come, ad esempio, la registrazione di un cliente, la produzione di un prodotto, l'acquisizione di un ordine, ecc. I dati mediati e raccolti dal processo sono altamente strutturati e includono transazioni, tabelle di riferimento e relazioni, nonché i metadati che definiscono il contesto. I dati aziendali tradizionali rappresentano la stragrande maggioranza di ciò che l'IT gestisce ed elabora, sia nei sistemi operativi che in quelli di business intelligence. Questa tipologia di dati è solitamente strutturata e viene archiviata in sistemi di database relazionali. Alcuni esempi di tali dati sono i seguenti:

- Dati prodotti dalla pubblica amministazione come, ad esempio, i dati sanitari
- Dati prodotti in ambito di business:
    - Transazioni commerciali
    - Dati finanziari, economici ed aziendali
    - Dati da E-commerce
    - Dati da carte di credito

3. **Internet of Things (dati generati da macchine).** Questi sono i dati che sono stati generati e si sono sviluppati in seguito alla crescita esponenziale del numero di sensori e macchine utilizzate per misurare e registrare gli eventi e le situazioni nel mondo fisico. I dati generati da questi sensori sono semplici record o anche complessi registri di computer, e quindi essi sono ben strutturati. Con la proliferazione dei suddetti sensori, il volume di questi dati aumenta a dismisura. Tali informazioni diventano una componente sempre più importante della massa di dati archiviati ed elaborati da molte aziende. La loro natura, ben strutturata, è molto adatta all'elaborazione informatica, anche se le loro dimensioni e la velocità di produzione vanno oltre gli approcci tradizionali. Alcuni esempi di tali dati sono i seguenti

- Dati da sensori:
    - Sensori fissi: automazione delle case, sensori meteorologici o di misurazione dei livelli di inquinamento, sensori o webcam per monitorare il traffico, sensori scientifici, sensori di videosorveglianza o di sicurezza
    - Sensori mobili: mobile phone location, sensori di automobili, immagini satellitari
- Dati da computer systems: logs e web logs

5.4 LA STATISTICA E I BIG DATA

Un mondo e un'economia così data-driven implicano come conseguenza che la gestione dei dati non è più un solo monopolio degli istituti di statistica o delle organizzazioni internazionali che gestiscono i dati ufficiali. Per la natura stessa dei big dati, i problemi e le domande di ricerca che si indagano richiedono team multidisciplinari formati da esperti di area tematica (di dominio), esperti di calcolo matematico, esperti di statistica e machine learning.
Durante un intervento alla *Johns Hopkins School of Public Health*, Roger Peng afferma che "*in Big data, statistical sciences and domain sciences are more intertwined than ever before, and statistical methodology is absolutely critical to making inferences*".
La statistica è quindi una scienza fondamentale per assicurare che le informazioni estratte da grandi insiemi di

dati siano significative ed accurate. In particolare, la statistica viene utilizzata per analizzare i seguenti problemi:

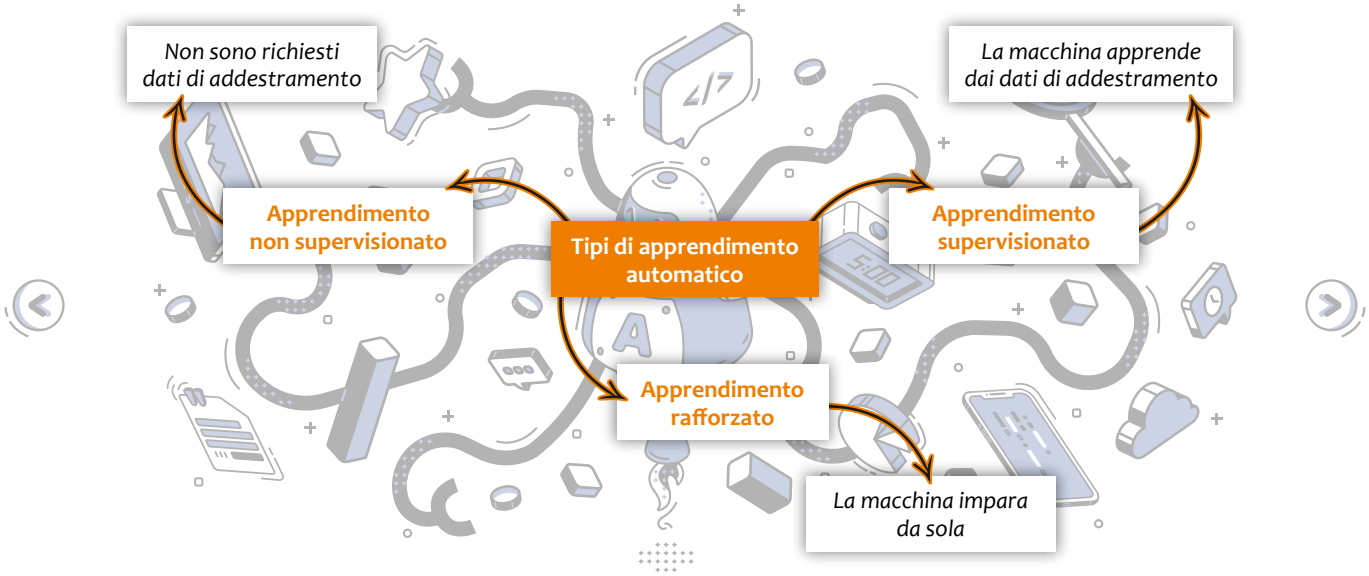
- Valutazione della qualità dei dati e imputazione dei dati mancanti
- Valutazione della natura dei dati osservati ed eventuali *counfounders*
- Metodi di quantificazione dell'incertezza nelle predizioni (*prediction*), nelle previsioni (*forecast*) e nei diversi modelli utilizzati

La statistica è in grado di mettere in campo tecniche e modelli sofisticati per fornire risposte adeguate a molteplici domande di ricerca, ponendo particolare attenzione a tre elementi fondamentali:

- la struttura dei dati
- il processo ("*data generating process*") sottostante che ha generato i dati, ossia l'identificazione del modello statistico
- il parametro o i parametri specifici del problema in esame che si desidera stimare o prevedere con i dati di interesse

5.5 METODI DI MACHINE E STATISTICAL LEARNING

Le tecniche di *Machine e Statistical learning* hanno l'obiettivo di comprendere la struttura dei dati e di associare ai dati in esame modelli che spiegano la dinamica e le interconnessioni osservate e che successivamente possono essere applicati in altre prove. Una macchina può imparare dai dati in diversi modi: in alcuni casi le macchine vengono addestrate adattando i dati a modelli statistici specifici, mentre in altri casi le macchine apprendono in maniera autonoma e identificano una struttura complessa sulla base di alcune indicazioni di ricerca. Per questo motivo tali metodi si differenziano in tecniche di *supervised learning* e in tecniche di *unsupervised learning*. A completamento, si ricorda che esistono anche tecniche di *reinforcement learning:* esse non si basano su modelli statistici ma utilizzano un processo del tipo "*hit and trial*", ossia la macchina impara dall'esperienza. Quando la prova ha successo, si ottiene una ricompensa; quando invece è fallimentare si ottiene una penalità, con l'obiettivo di massimizzare la ricompensa totale.



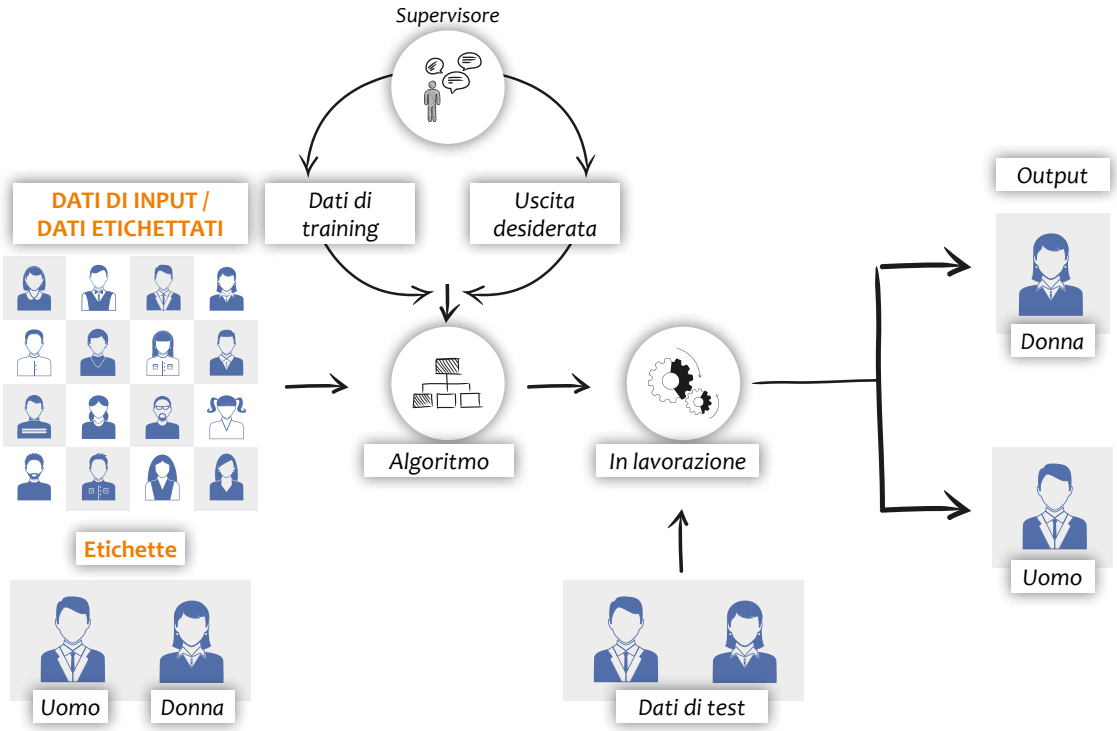
- Nota: le piattaforme di social media come Facebook, Instagram, LinkedIn, ecc. utilizzano algoritmi di machine learning per la classificazione degli utenti. Per esempio, Facebook nota e registra le tue attività, le tue chat, i tuoi "like", i commenti e il tempo trascorso su specifici tipi di post, video, ecc. L'algoritmo poi impara le attività e le esperienze preferite dell'utente e, a sua volta, suggerisce amicizie o crea suggerimenti di pagine per tale profilo.

TECNICHE DI SUPERVISED LEARNING

Gli algoritmi di *supervised learning* sono progettati per imparare la relazione dei dati dai modelli che lo studioso identifica e capire la struttura dei dati identificata tramite esempi. Infatti, questo metodo viene chiamato apprendimento supervisionato perché l'algoritmo apprende dal set di dati di addestramento (*training set*), esattamente come un insegnante supervisiona il processo di apprendimento dei propri studenti. In una prima fase, ai dati viene assegnata un'etichetta (*label*) e vengono identificate le variabili di input e le variabili di output. Durante il processo di apprendimento, l'algoritmo identifica la struttura nei dati di input in relazione ai dati di output, ossia esso è in grado di stimare i parametri ignoti che caratterizzano il modello imposto dai dati. Dopo aver completato la fase di training, l'algoritmo osserverà nuovi dati di input e determinerà quale etichetta assegnare al nuovo elemento in base alla classificazione dei dati svolta durante la fase di addestramento. In sintesi, l'obiettivo degli algoritmi di *supervised learning* è quello di prevedere la corretta assegnazione dei nuovi dati di input che entrano nel sistema sulla base del modello stimato in precedenza.

Supponiamo di avere un dataset di immagini di uomini e donne. Ad ogni immagine del set di dati viene asse-

gnata un'etichetta. Questo significa che sappiamo riconoscere quale immagine rappresenta un uomo e quale immagine rappresenta una donna. Nella fase di addestramento, tutte le immagini e le etichette corrispondenti vengono elaborate dall'algoritmo che sarà quindi in grado di differenziare le immagini. Se l'algoritmo ha imparato bene dalla fase di training, quando verrà mostrata una nuova immagine, questa volta senza etichetta, esso dovrebbe essere in grado di stabilire se tale immagine viene associata ad un uomo o ad una donna.



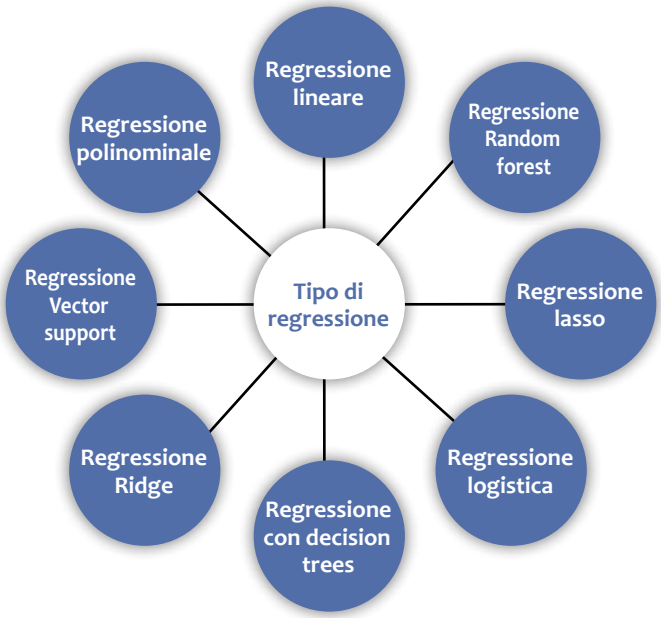
Le tecniche di *supervised learning* si dividono in problemi di regressione e classificazione.

Con riferimento al primo tipo di procedure, gli algoritmi di regressione vengono utilizzati se esiste una relazione tra una o più variabili di input e le variabili di output. L'obiettivo è quello di prevedere un valore molto vicino al valore reale di uscita e quindi la valutazione del successo viene perseguita con il calcolo dell'errore commesso nella previsione. Più piccolo è l'errore, maggiore è la precisione del modello di regressione. Gli algoritmi di regressione vengono utilizzati se esiste una relazione tra le variabili di input e la variabile di output, quando il valore della variabile di output è continuo o reale, come, ad esempio, il prezzo della casa, il prezzo delle azioni o le vendite di un esercizio commerciale. La seguente tabella mostra un set di dati che serve per prevedere le vendite in euro in una catena di negozi, sulla base di diversi parametri. Qui, le variabili di input sono il numero di spot pubblicitari realizzati nel mese e il numero di addetti alle vendite nei diversi punti vendita mentre la variabile di output è il valore (continuo) delle vendite in euro. L'obiettivo di questo problema di regressione consiste nel prevedere un valore molto vicino al valore reale di output sulla base della stima del modello.

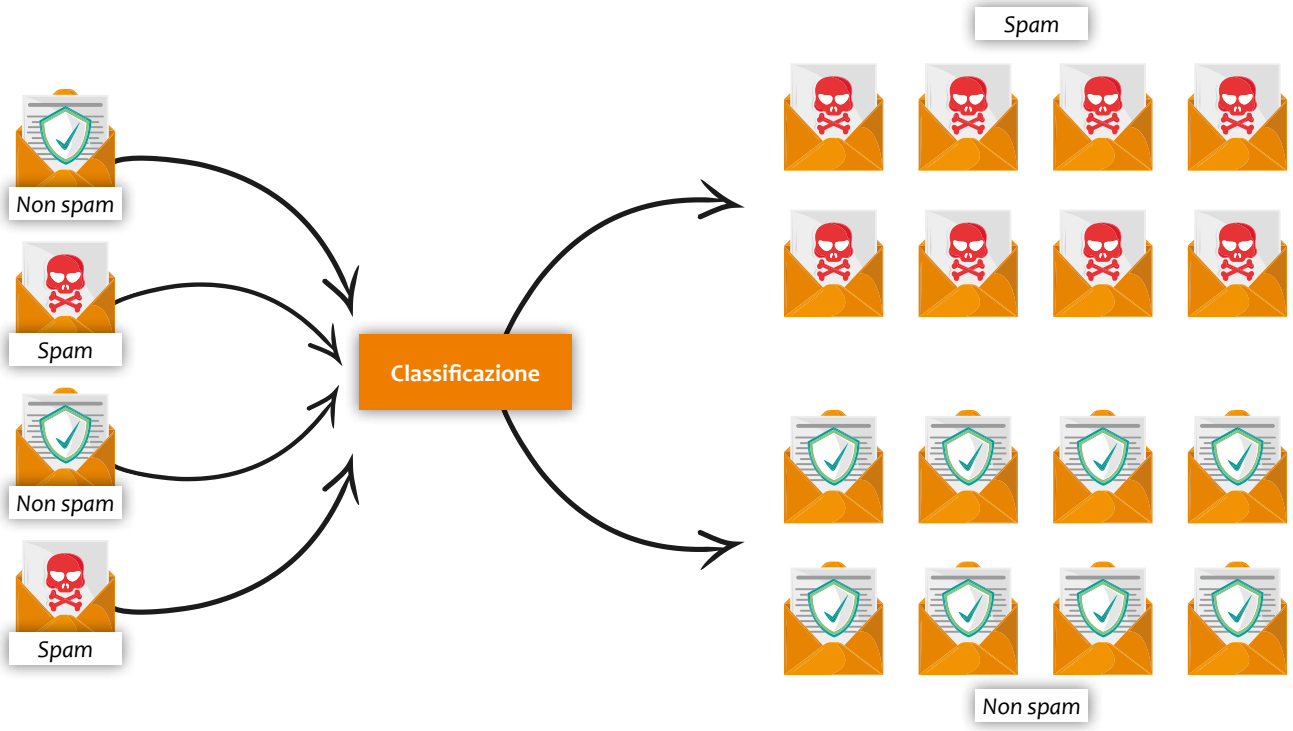
Numero di spot pubblicitari nel mese	Numero di addetti alle vendite	Vendite in euro
45	130	97
47	128	95
40	135	94
36	119	92
35	124	90

37	120	85
32	117	83
30	112	76
25	115	73
27	108	71

I metodi di regressione sono molteplici e al solo scopo illustrativo ne riportiamo una panoramica in figura.

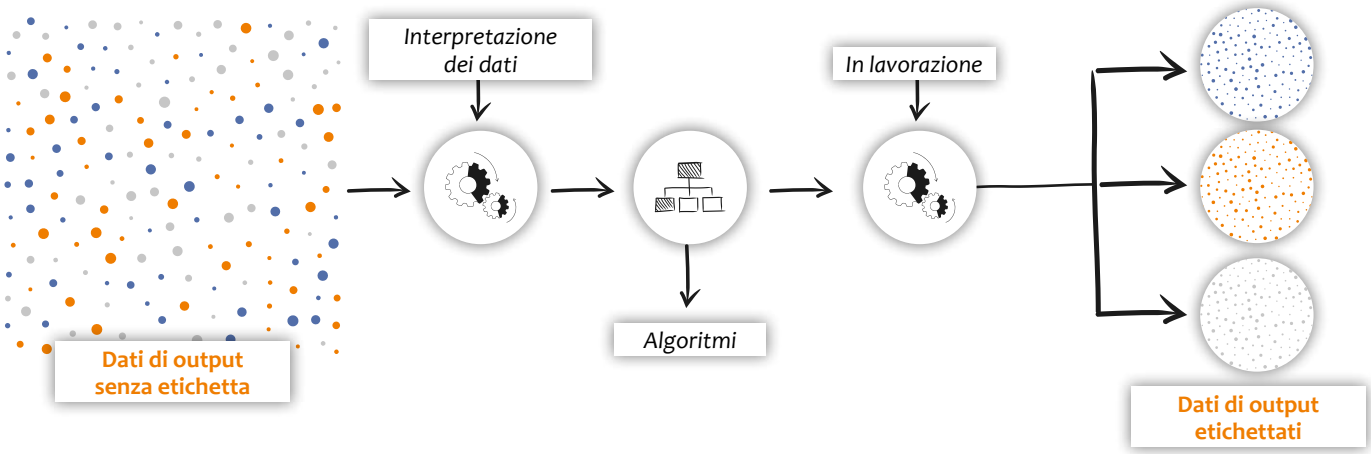


Il secondo gruppo di *supervised learning* si riferisce agli algoritmi di classificazione che mirano a raggruppare l'output in diverse classi sulla base di più variabili di input. Solitamente si utilizzano tecniche di classificazione quando il valore della variabile di output è discreto o in categorie. Ad esempio, un simile processo accade ogni giorno nella vostra casella mail quando il sistema classifica le mail in spam e non spam.



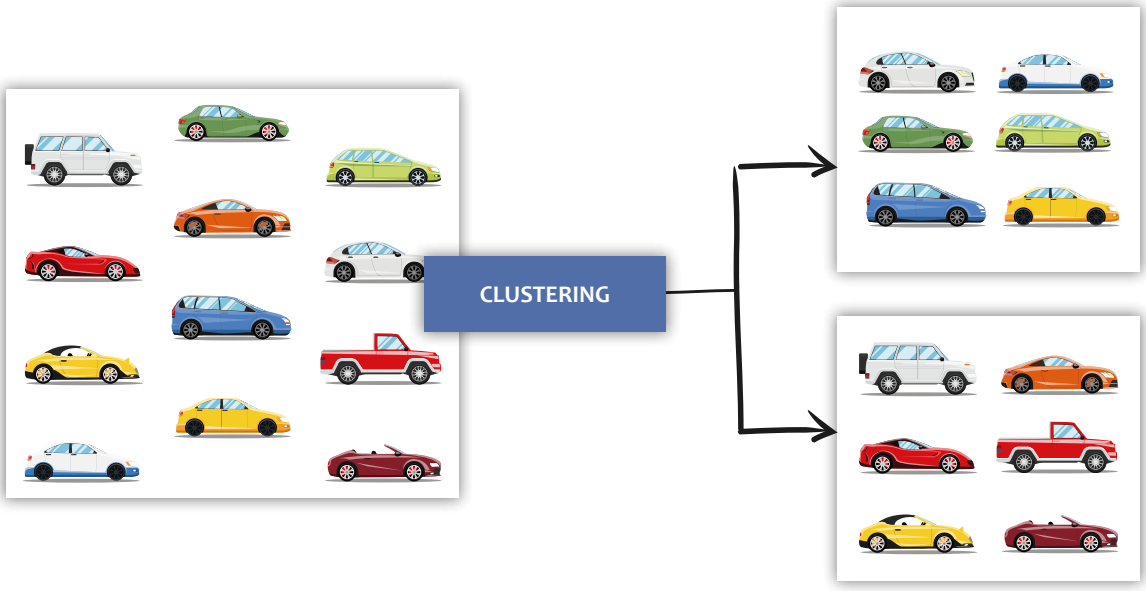
TECNICHE DI UNSUPERVISED LEARNING

Negli algoritmi di *unsupervised learning*, i dati non vengono etichettati e quindi non si osservano dati di input e nemmeno variabili di output corrispondenti. Gli utenti non devono addestrare o supervisionare il modello; infatti, non esiste un output corretto a priori. L'algoritmo stesso apprende la struttura e le relazioni dai dati e ne identifica le informazioni per raggruppare i dati secondo le proprie somiglianze. Nel seguente esempio l'algoritmo identifica i punti di diverso colore e li classifica in diversi gruppi.



Gli algoritmi di *unsupervised learning* si dividono in problemi di clustering e di associazione.

Le tecniche di clustering hanno l'obiettivo di fornire strutture a dati non strutturati. Il processo di raggruppamento consiste nel dividere i punti del campione di dati o dell'intera popolazione in alcuni gruppi (in numero molto inferiore alla dimensione dei dati) in modo tale che i punti appartenenti allo stesso gruppo hanno caratteristiche simili, mentre quelli situati in gruppi diversi hanno proprietà differenti. Questi gruppi sono chiamati *cluster*. Un semplice esempio di clustering è mostrato nella figura sottostante, dove diverse automobili vengono raggruppate in due cluster differenziando tra auto utilitarie e auto sportive. L'utente può altresì scegliere il numero di cluster modificando di conseguenza la granularità dei gruppi.

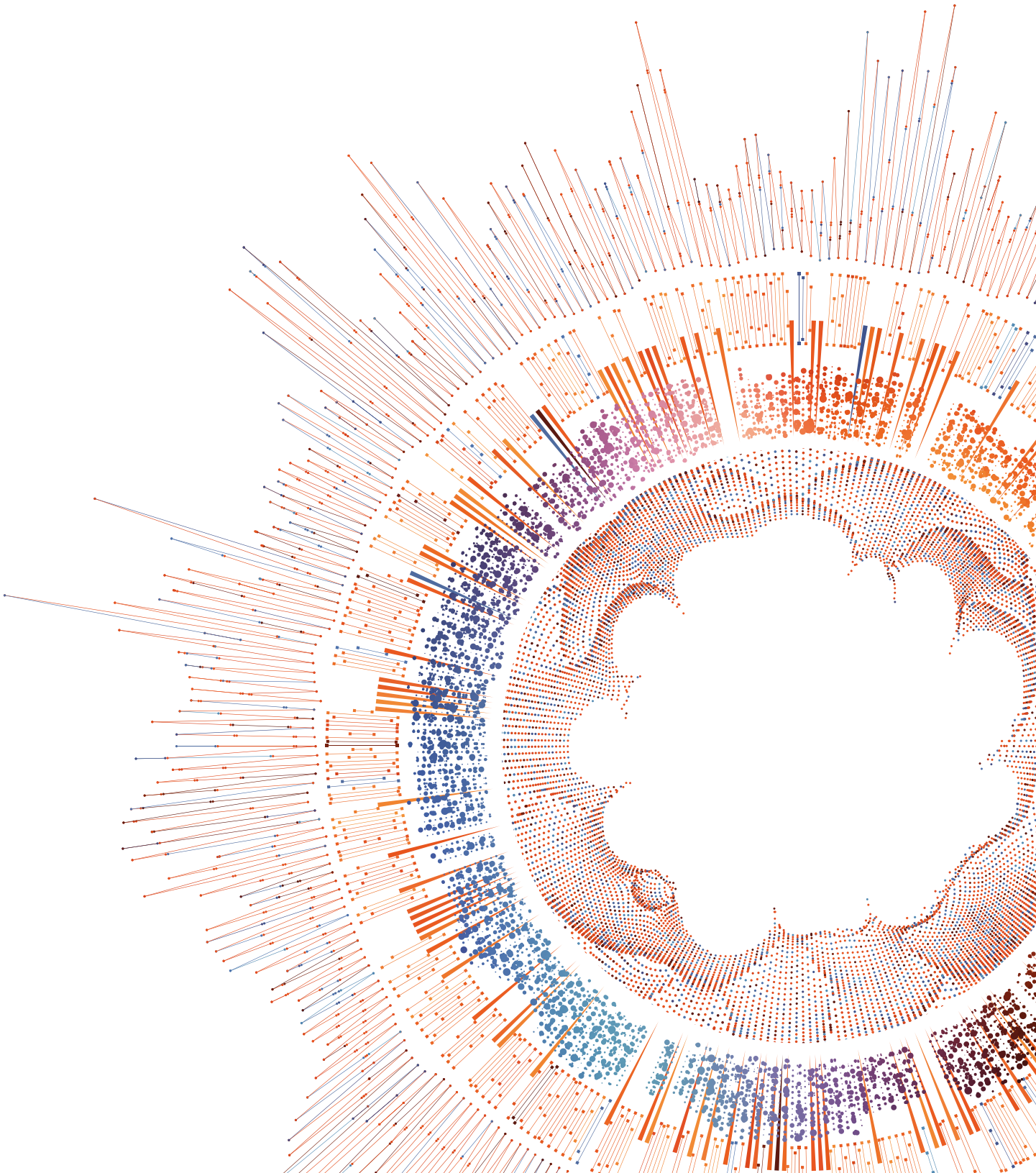


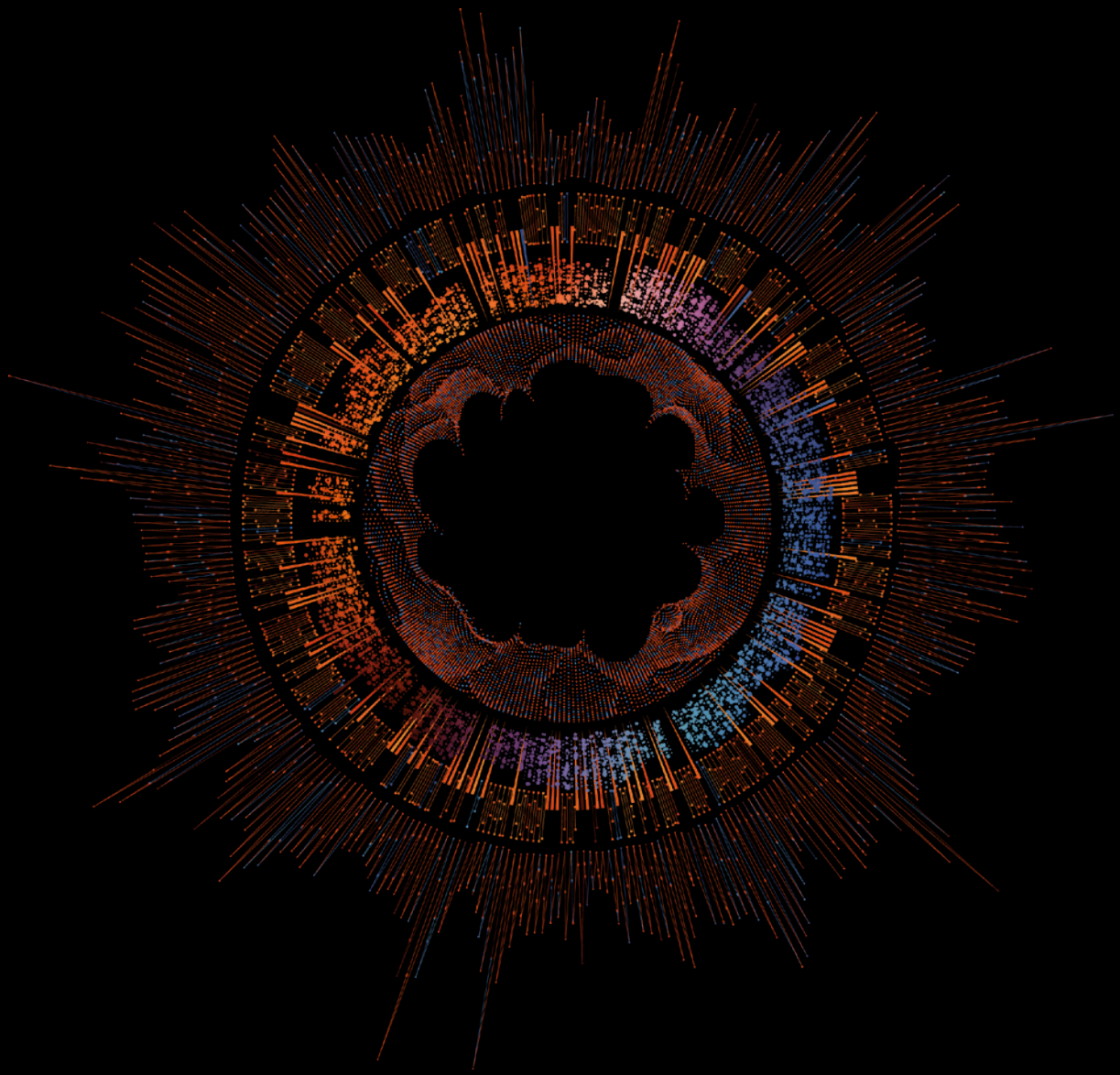
Le tecniche di associazione sono invece utilizzate per trovare la relazione, l'associazione, e/o le dipendenze tra gli elementi presenti in un dataset. Questi algoritmi utilizzano varie regole decisionali per determinare relazioni utili nei dati in esame. L'apprendimento delle regole di associazione è particolarmente vantaggioso per le organizzazioni aziendali che sviluppano prodotti. Infatti, la mappatura delle relazioni tra i diversi prodotti di queste organizzazioni possono migliorare le loro vendite e aumentare i profitti.
Un classico esempio riguarda la m*arket basket analysis*. Questa osserva infatti la strategia di esposizione dei prodotti in un supermercato. Tutti i prodotti che possono essere acquistati insieme vengono messi nello stesso scaffale o nelle loro vicinanze. Per esempio, se un cliente desidera acquistare pane, probabilmente egli acquisterà anche burro, uova o latte; per questo motivo, questi articoli vengono spesso posizionati su scaffali nelle immediate vicinanze.

AL LAVORO!

1. SIA LE TECNICHE DI SUPERVISED LEARNING CHE QUELLE DI UNSUPERVISED LEARNING NECESSITANO DI...
   A. UN ATTRIBUTO IN CATEGORIE
   B. UN OUTPUT
   C. UN INPUT

2. QUALE DEI SEGUENTI È UN COMPITO PER UN ALGORITMO DI UNSUPERVISED LEARNING?
   A. RAGGRUPPARE IMMAGINI DI SCARPE E CAPPOTTI SEPARATAMENTE PER UN DETERMINATO SET DI IMMAGINI
   B. IMPARARE A GIOCARE A SCACCHI
   C. PREVEDERE SE UN ALIMENTO È DOLCE O PICCANTE IN BASE AD INFORMAZIONI SUGLI INGREDIENTI E SULLE LORO QUANTITÀ





ISBN: 978-88-89427-04-0

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA