Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
SCIENZE STATISTICHE

RELAZIONE FINALE

"FACCIAMO STATISTICA A SCUOLA": MODULI DIDATTICI PER L'INSEGNAMENTO DELLA STATISTICA NELLE SCUOLE SECONDARIE DI SECONDO GRADO

Relatore: Prof.ssa Laura Ventura
Dipartimento di Scienze Statistiche

Laureando: Matteo Gollin
Matricola: 1114044

Anno Accademico 2017/2018

"Proveranno a dirti chi sei per tutta la vita. Tu reclama il tuo spazio e dimostra chi sei davvero. Se vuoi che ti guardino diversamente, datti da fare. Se vuoi che le cose cambino, devi riuscire a cambiarle tu stesso, perché al mondo non esistono Fate Madrine."

JENNIFER MORRISON - Emma Swan

Indice

Introduzione

Da sempre una delle discipline fondamentali nella formazione alla cittadinanza è la Matematica. Nasce, quindi, la necessità di costruire una cultura comune in questa materia che rafforzi conoscenze e competenze nella formazione degli studenti che andranno a completare la scuola secondaria di II grado.
Inoltre c'è da tenere in considerazione la crescente importanza della Statistica nel mondo del lavoro, con la conseguente esigenza di introdurre tale disciplina nelle scuole per avvicinare gli studenti a questa scienza. Il problema risiede nell'accorpamento delle due classi di concorso relative a Matematica e Statistica a fronte di un ridimensionamento orario previsto dai decreti D.P.R. 87, 88, 89 del 15 maggio 2010.
A causa di questo riordino stabilito dal Ministero, gli insegnanti vedono aumentati gli obiettivi formativi ma diminuite le ore a disposizione per conseguirli in maniera adeguata. Di conseguenza, per riuscire a completare nell'anno scolastico il programma designato, i docenti sono costretti a rimuovere determinati argomenti dall'insegnamento. Diverse sono le scelte in merito a quali di questi eliminare ma, in generale, le decisioni ricadono su quelli inerenti alla Statistica. Perciò risulta necessario aiutare gli insegnanti a prendere maggior consapevolezza dell'importanza delle scienze statistiche e facilitare nella loro professione l'insegnamento di tale disciplina. Questo rappresenta uno degli scopi del progetto Piano Lauree Scientifiche (PLS), che si pone anche l'obiettivo di far conoscere agli studenti una materia sconosciuta ed orientarli per un'eventuale scelta universitaria.
Lo scopo di questa tesi consiste nel creare moduli didattici, inerenti all'insegnamento della Statistica, da proporre ai docenti delle scuole superiori da

portare nelle classi. L'intenzione è quella di aiutarli ad ottimizzare le risorse ed avvicinare gli studenti al mondo della Statistica, nel miglior modo possibile. Questo perché i docenti che possiedono un'impronta matematica, tendono ad insegnare la Statistica attraverso formule, abbandonando la vera sostanza che contraddistingue questa disciplina.
I moduli proposti hanno lo scopo di catturare l'interesse dello studente, con argomenti che risultano moderni ed intriganti ma, allo stesso tempo, inerenti a ciò che si appresta a studiare durante il percorso scolastico. Il limite orario imposto dal riordino, permette l'insegnamento di pochi concetti semplici agli studenti delle scuole superiori e, in particolare, inerenti alla statistica descrittiva. Pertanto i moduli didattici vengono costruiti sulla base di questa concezione, limitandosi a far vedere applicazioni semplici su fenomeni reali. Questa tesi si presta, peciò, ad una duplice lettura. Da un lato è necessario affrontare contenuti intuitivi per i ragazzi delle scuole superiori, che possiedono conoscenze basilari inerenti alla Statistica. Dall'altro, sul piano statistico, è importante dedicare il giusto peso alla parte tecnica ed applicativa.
Un altro obiettivo molto importante del progetto PLS è dimostrare ai ragazzi l'interdisciplinarità della Statistica, attraverso l'applicazione di strumenti statistici per l'analisi di dati reali, a prescindere dall'ambito su cui sono stati rilevati.
La tesi è strutturata nel seguente modo:

- Nel *Capitolo 1* si presenta la situazione attuale dell'insegnamento della Matematica nelle scuole secondarie di II grado e le conseguenze del riordino decretato dal Ministero dell'Istruzione, tra cui l'inserimento della Statistica.

- Nel *Capitolo 2* si parla dell'importanza della Statistica nella scuola e si tracciano delle linee guida per indirizzare i docenti nell'insegnamento della Statistica in classe.

- Nel *Capitolo 3* viene descritto un primo modulo didattico per l'insegnamento della Statistica che tratta il tema della *correlazione*, in cui viene analizzato un *dataset* inerente alle acque potabili del Veneto per verificare un'ipotetica correlazione positiva tra Ferro e Ammonio.

- Nel *Capitolo 4* viene illustrato un secondo modulo didattico che descrive la *Sentiment Analysis* applicata a dati tratti da Twitter per trovare la distribuzione aggregata delle opinioni delle persone in riferimento alla soddisfazione di tre marche di cellulari.

- Nel *Capitolo 5* si trattano le conclusioni e, in particolare, l'esperienza con gli studenti dell'Istituto "E. Fermi" di Bassano del Grappa, svoltosi in data 04 Dicembre 2017 all'interno del PLS del Dipartimento di Scienze Statistiche dell'Università degli Studi di Padova (`https://pls.scienze.unipd.it/statistica/`).

Capitolo 1

La Matematica nel riordino della Scuola secondaria di II grado

In questo capitolo si fornisce una descrizione dello stato attuale dell'insegnamento della Matematica nelle scuole secondarie di secondo grado. Per comprendere appieno la situazione si porge particolare attenzione al ridimensionamento del quadro orario e alle Indicazioni Nazionali per i Licei, nonché alle Linee guida per gli Istituti Tecnici e Professionali (http://archivio.pubblica.istruzione.it/riforma_superiori/nuovesuperiori/index.html).

1.1 I quadri orari nel riordino

Nei decreti D.P.R. 87, 88, 89 del 15 maggio 2010 emanati dal Ministero dell'Istruzione (MIUR) si parla di *riordino* della scuola secondaria di II grado. Le tematiche principali riguardono, prevalentemente, gli obiettivi formativi delle discipline scolastiche e il ridimensionamento orario settimanale delle stesse.
Nello specifico, la situazione attuale dei quadri orari dedicati alla matematica è riportata nella Tabella 1.1.

	I	II	III	IV	V
Liceo Classico (Linguistico, Scienze Umane)	3	3	2	2	2
Liceo Scienze Umane (opzione economico-sociale)	3	3	3	3	3
Liceo Scientifico	5	5	4	4	4
Liceo Scientifico (opzione scienze applicate)	5	4	4	4	4
Istituti Tecnici e Professionali	4	4	3(+1)	3(+1)	3(+1)

Tabella 1.1: Quadri orari nel riordino.

A differenza della situazione precedente, nel riordino si è osservata una diminuzione in media degli orari settimanali dedicati all'insegnamento della Matematica in quasi tutti gli indirizzi di studio. L'unica nota positiva riguarda i Licei Scientifici, dove si è potuto notare un aumento di ore settimanali, passando dalla combinazione 5-4-3-3-3 a quella definita nella Tabella 1.1.
In confronto ad altre materie che hanno subito una riduzione di orario, la Matematica risulta una delle discipline più penalizzate se si ragiona in termini di obiettivi formativi e tipologie di scuola. Si pensi, ad esempio, ad un paragone con il Latino nei Licei, dove si può riscontrare per entrambe le materie un ridimensionamento orario. Risulta necessario fare un confronto tra il numero di studenti del Liceo Classico che usufruiranno in futuro della Matematica e gli studenti del Liceo Scientifico che utilizzeranno il Latino. In proporzione, la riduzione di orario applicata all'insegnamento della Matematica dovrebbe risultare inferiore rispetto a quella impartita al Latino.
Per quanto riguarda gli Istituti Tecnici, nel secondo triennio e quinto anno si è riscontrata una drastica diminuzione delle ore dedicate alla Matematica, nonché una comunanza in termini di indirizzi di studio. Per esempio, prima del riordino erano previste 6 ore settimanali per l'indirizzo informatico degli Istituti Tecnici e 4 per i restanti indirizzi. Con la riforma sono state fissate 3 ore, indipendentemente dall'indirizzo di studio intrapreso.
Tuttavia è stata prevista un'ora aggiuntiva settimanale dedicata all'attività *Complementi di Matematica*. Questo insegnamento prevede diverse attività

specifiche per ogni indirizzo di studio, non necessariamente impartite dal docente di Matematica. In molti casi, quest'ora di complementi viene utilizzata per svolgere moduli didattici di impronta statistica, a discrezione dell'insegnante.
Oltre al ridimensionamento orario delle discipline scolastiche, il riordino prevede anche indicazioni di carattere metodologico. Infatti sono state pubblicate le *Indicazioni nazionali* per i Licei e le *Linee guida* per gli Istituti Tecnici e professionali. Tra le due stesure sussiste una sostanziale differenza, ma è opinione diffusa che sarebbe risultata più efficiente l'elaborazione di un'unica struttura.

1.2 Le Indicazioni Nazionali per i Licei

Le Indicazioni Nazionali pensate per i Licei, a prescindere dagli indirizzi di studio, prevedono una ripartizione delle conoscenze nei seguenti ambiti:

- *Aritmetica e algebra*
- *Geometria*
- *Relazioni e funzioni*
- *Dati e previsioni*

I primi due settori hanno subito una variazione in termini di dicitura rispetto al curricolo ministeriale *Matematica 2003*, nel quale i relativi nomi erano *Numeri e algoritmi* e *Spazio e figure*. Mentre per quest'ultimi si è pensato ad un cambiamento, optando per dei nomi più tradizionali, per gli altri due ambiti si è deciso di conservare gli stessi titoli assegnati dal decreto.
Di particolare importanza, per questa tesi, è il nucleo *Dati e previsioni* in quanto caratterizzato da un'impronta prettamente statistica.
Per ogni settore è necessario specificare con chiarezza il quadro generale di riferimento dei concetti fondamentali e gli obiettivi formativi, in quanto le scuole e gli insegnanti hanno bisogno di conoscere con precisione quali sono. Prima di vedere nel dettaglio il quadro di riferimento degli apprendimenti, è

fondamentale precisare che la suddivisione del profilo generale dei licei risulta essere la seguente:

- *Primo biennio*
- *Secondo biennio*
- *Quinto anno*

Per ogni profilo è stata tracciata una linea guida sugli argomenti principali, per ciascun nucleo di apprendimento. Lo scopo è fornire un punto di riferimento per coloro che si occupano della didattica della Matematica nelle scuole.

1.2.1 La Matematica nei Licei: Primo Biennio

Tra i quattro settori citati precedentemente, il più rilevante per questo progetto è quello inerente alla Statistica (*Dati e previsioni*). Ciò nonostante non si può distogliere l'attenzione dagli altri tre ambiti, che rappresentano il blocco principale dell'insegnamento della Matematica in tutte le tipologie di scuola. Perciò risulta necessario fare un excursus sulle Indicazioni Nazionali relative a questi nuclei, nello specifico nel primo biennio dei Licei.

Aritmetica e algebra
Concetti di vettore, di dipendenza lineare, di prodotto scalare e vettoriale nel piano e nello spazio; elementi del calcolo matriciale.

Geometria
Funzioni circolari e le loro proprietà e relazioni elementari; i teoremi che permettono la risoluzione dei triangoli e il loro uso nell'ambito di altre discipline, in particolare nella fisica.
Studio delle funzioni quadratiche, rappresentazione geometrica delle coniche. Trasformazioni geometriche (traslazioni, rotazioni, simmetrie, similitudini), proprietà invarianti.

Come si può notare, argomenti inerenti all'ambito **Relazioni e funzioni** non sono previsti per i primi due anni scolastici.
Osservando le linee guida si può arrivare alla conclusione che il primo biennio dei Licei sia troppo carico di argomenti, a fronte del quadro orario previsto dal riordino. Quindi se da un lato la proposta può risultare condivisibile dal punto di vista metodologico, dall'altro diventa problematica la collocazione degli argomenti nelle prime due classi. Pertanto sarebbe consigliabile inserire alcuni contenuti prevalentemente nel II biennio.

Per quanto riguarda il settore **Dati e previsioni** le normative vigenti prevedono che lo studente apprenda le seguenti nozioni:

- *nozione di probabilità, con esempi tratti da contesti classici e con l'introduzione di nozioni di statistica;*
- *concetto di modello matematico, distinguendo la specificità concettuale e metodica rispetto all'approccio della fisica classica.*

Una delle novità più importanti inserite nelle Indicazioni Nazionali è l'introduzione di argomenti riguardanti la Statistica e la Probabilità in tutti gli indirizzi di studio. Ciò comporta una diversa impostazione della didattica e un maggior collegamento con le applicazioni e il mondo reale. Prima del riordino, argomenti di questa natura erano presenti solo in corsi sperimentali (es. indirizzo informatica industriale negli Istituti Tecnici). Tuttavia queste tematiche erano molto trascurate dagli insegnanti, in quanto la loro impronta matematica li portava a preferire argomenti inerenti agli altri tre ambiti di apprendimento.
La situazione attuale, nonostante la riforma, rimane invariata per la maggior parte delle istituzioni scolastiche.

1.2.2 La Matematica nei Licei: Secondo Biennio

Nel secondo biennio dei Licei le tematiche che vengono affrontate nell'insegnamento della Matematica riguardano, principalmente, gli ambiti di *Rela-*

zioni e funzioni e *Dati e previsioni*. Per quanto riguarda il primo settore, le Indicazioni Nazionali si presentano particolarmente interessanti dal punto di vista didattico, introducendo nozioni di analisi matematica utili per coloro intendono proseguire il loro percorso all'università.

Relazioni e funzioni
Lo studente apprende ad analizzare sia graficamente che analiticamente le principali funzioni e impara ad operare su funzioni composte e inverse. Un tema importante di studio è il concetto di velocità di variazione di un processo rappresentato mediante una funzione, con riferimento al rapporto incrementale e allo studio degli incrementi di una funzione "a passo costante" come introduzione al concetto di derivata.

Il nucleo *Dati e previsioni* assume un ruolo più rilevante in questo secondo biennio, introducendo tematiche fondamentali nella concezione statistica.

Dati e previsioni
Lo studente, in ambiti via via più complessi, apprende a far uso delle distribuzioni doppie condizionate e marginali, dei concetti di deviazione standard, dipendenza, correlazione e regressione, di campione. Studia la probabilità condizionata e composta, la formula di Bayes e le sue applicazioni, nonché gli elementi di base del calcolo combinatorio. Lo studio deve essere sviluppato il più possibile in collegamento con le altre discipline, con la possibilità che i dati vengano raccolti direttamente dagli studenti.

Nonostante i contenuti siano importanti, si riscontra un'incongruenza in merito all'insegnamento del calcolo combinatorio. Infatti le indicazioni prevedono di insegnare tale argomento dopo aver introdotto nozioni di calcolo delle probabilità. Pertanto sarebbe opportuno anticipare i concetti più semplici del calcolo combinatorio al I biennio per riuscire a svolgere in maniera adeguata i contesti classici della probabilità nel II.

1.2.3 La Matematica nei Licei: Quinto Anno

Nel quinto anno lo studente si appresta a terminare il suo percorso di studio nella scuola secondaria, concentrandosi nella preparazione all'esame di stato e all'ingresso nel mondo universitario. Perciò risulta necessaria la produzione di un syllabus delle conoscenze e delle abilità che possono essere soggette alla valutazione nell'esame finale. Inoltre serve un accordo preciso con le Università, specialmente con le Facoltà scientifiche, per comprendere quali siano le abilità richieste e i concetti che si aspettano di trovare nella preparazione dello studente.

Relazioni e funzioni
Lo studente prosegue lo studio delle funzioni fondamentali dell'analisi anche attraverso esempi tratti dalla fisica o da altre discipline. Acquisisce il concetto di limite di una successione e di una funzione e apprende a calcolare i limiti in casi semplici.
Lo studente acquisisce i principali concetti del calcolo infinitesimale - in particolare la continuità, la derivabilità e l'integrabilità - anche in relazione con altre problematiche (velocità istantanea in meccanica, tangente di una curva, calcolo di aree e volumi).
Altro importante tema di studio è il concetto di equazione differenziale e principali proprietà, con particolare attenzione per l'equazione della dinamica di Newton.

Geometria
Lo studente apprende i primi elementi di geometria analitica dello spazio e la rappresentazione analitica di rette, piani e sfere.

Oltretutto, nell'anno finale lo studente approfondisce la comprensione del metodo assiomatico e la sua utilità concettuale e metodologica dal punto di vista della modellazione matematica. Gli esempi vengono tratti dal contesto dell'aritmetica, della geometria euclidea o della **probabilità**. Ciò nonostante è lasciata all'insegnante la decisione di quale settore disciplinare privilegiare

allo scopo. Purtroppo, ancora una volta, l'ambito statistico viene penalizzato data la formazione prevalentemente matematica dei docenti.

1.3 Le Linee guida per gli Istituti Tecnici e Professionali

Le Linee guida per gli Istituti Tecnici e Professionali sono state redatte in maniera differente rispetto alle Indicazioni Nazionali per i Licei. Infatti presentano una struttura a due colonne in cui vengono definite le *conoscenze* in una e le *abilità* nell'altra. Tuttavia non sussistono differenze significative tra le linee degli Istituti Tecnici e quelle degli Istituti Professionali.
La suddivisione dei settori disciplinari rimane la stessa delle Indicazioni Nazionali per i Licei:

- *Aritmetica e algebra*
- *Geometria*
- *Relazioni e funzioni*
- *Dati e previsioni*

Il docente di Matematica ha lo scopo di far conseguire allo studente, al termine del percorso quinquennale, risultati di apprendimento che lo mettano in grado di padroneggiare diversi concetti matematici, anche in ambito statistico. In particolare deve possedere gli strumenti statistici e del calcolo delle probabilità necessari per la comprensione delle discipline scientifiche, nonché per poter operare nel campo delle scienze applicate.
Per quanto riguarda la ripartizione del profilo scolastico, anche gli Istituti Tecnici e Professionali sono caratterizzati dalla divisione in *Primo Biennio*, *Secondo Biennio* e *Quinto anno*.

1.3.1 Le Linee guida per il Primo Biennio

Per ciascun nucleo di apprendimento vengono definite le *Conoscenze* e le *Abilità* che lo studente deve conseguire nel Primo Biennio scolastico.

Per quanto concerne il settore **Aritmetica e algebra**, gli argomenti principali previsti dalle Linee guida sono i seguenti:

- Insiemi numerici: numeri naturali, interi, razionali (sotto forma frazionaria e decimale), irrazionali e reali; ordinamento e loro rappresentazione su una retta; le operazioni con i numeri interi e razionali e loro proprietà.
- Potenze e radici; rapporti e percentuali; approssimazioni.
- Le espressioni letterali e i polinomi; operazioni con i polinomi.

Le abilità richieste riguardano l'utilizzo delle procedure di calcolo aritmetico per calcolare espressioni e risolvere problemi, nonché la capacità di operare con potenze e radicali utilizzando correttamente il concetto di approssimazione. Infine saper padroneggiare l'uso della lettera come variabile ed eseguire operazioni con i polinomi (tra cui la fattorizzazione).

Le conoscenze inerenti al nucleo di **Geometria** sono simili a quelle definite per i Licei e prevedono:

- Gli enti fondamentali della geometria (assioma, postulato, teorema e dimostrazione); nozioni fondamentali di geometria del piano e dello spazio; le principali figure del piano e dello spazio.
- Il piano euclideo: relazioni tra rette, congruenza di figure, poligoni e loro proprietà; circonferenza e cerchio; perimetro e area dei poligoni; teoremi di Euclide e Pitagora.
- Teorema di Talete; principali trasformazioni geometriche e loro invarianti (isometrie e similitudini).

Lo studente deve essere in grado di eseguire costruzioni geometriche elementari nonché conoscere e saper calcolare perimetro, area e volume delle principali figure geometriche del piano e dello spazio. Importante è la capacità di saper porre, analizzare e risolvere problemi del piano e dello spazio utilizzando le proprietà delle figure geometriche e/o isometrie.

A differenza del primo biennio dei Licei, negli Istituti Tecnici e Professionali si prevede l'insegnamento di tematiche relative al settore disciplinare **Relazioni e funzioni**, trattate in maniera meno approfondita con accezione più pratica che teorica.

- Funzioni di vario tipo (lineari, quadratiche, circolari,ecc.) e loro rappresentazione (numerica, funzionale e grafica); linguaggio degli insiemi e delle funzioni (dominio, composizione, inversa, ecc.)
- Equazioni e disequazioni di primo e secondo grado; sistemi di equazioni e disequazioni.
- Il metodo delle coordinate: il piano cartesiano; rappresentazione grafica delle funzioni.

Le abilità che si richiedono allo studente riguardano la risoluzione di equazioni, disequazioni e sistemi oltre che la rappresentazione sul piano cartesiano delle principali funzioni studiate. Inoltre è richiesta la capacità di risolvere problemi che implicano l'uso di questi strumenti, anche per via grafica, come primo passo verso la modellizzazione matematica.

Infine relativamente all'ambito **Dati e previsioni**, sono stati designati i seguenti obiettivi formativi:

- Dati, loro organizzazione e rappresentazione; distribuzioni delle frequenze a seconda del tipo di carattere e principali rappresentazioni grafiche; valori medi e misure di variabilità.
- Significato della probabilità e sue valutazioni; semplici spazi (discreti) di probabilità: eventi disgiunti, probabilità composta, eventi indipendenti; probabilità e frequenza.

Le Linee guida richiedono allo studente la capacità di raccogliere, organizzare e rappresentare un insieme di dati, oltre che a calcolare i valori medi e alcune misure di variabilità di una distribuzione. Per di più si richiede la capacità di calcolare la probabilità di eventi elementari.

In generale, gli argomenti trattati nel primo biennio sono in numero molto elevato rispetto al monte ore che è previsto per l'insegnamento della Matematica. I docenti si trovano di fronte ad una decisione: sacrificare la qualità della didattica per la quantità oppure porre maggior attenzione ad un numero ristretto di argomenti. La scelta migliore per gli studenti è la seconda, tuttavia è necessario trovare un trade-off tra le due situazioni.

1.3.2 Le Linee guida per il Secondo Biennio e il Quinto anno

Nel "Triennio" finale degli Istituti Tecnici e Professionali, gli argomenti trattati riguardano prevalentemente i settori *Relazioni e funzioni* e *Dati e previsioni*.
Per quanto riguarda il primo dei due nuclei si prevedono obiettivi formativi inerenti al piano cartesiano e allo studio di funzione. Le Linee guida tracciate, prevedono l'insegnamento dei seguenti argomenti:

- Il piano cartesiano e la retta; la parabola e sue proprietà; risoluzione di disequazioni di secondo grado (intere e frazionarie) col metodo grafico della parabola.

- Le funzioni: concetto di funzione e relativa rappresentazione grafica; le funzioni esponenziali e i logaritmi; goniometria: le funzioni goniometriche, la misura degli angoli, equazioni e disequazioni con funzioni goniometriche.

- Calcolo infinitesimale: i limiti, le funzioni continue e il calcolo dei limiti; la derivata di una funzione e i teoremi del calcolo differenziale; lo studio delle funzioni; differenziale di una funzione; gli integrali indefiniti e i metodi di integrazione; gli integrali definiti; calcolo di aree e volumi; integrali impropri ed integrazione numerica.

- Equazioni differenziali; serie numeriche; serie di funzioni e di potenze.

Le abilità richieste per gli Istituti Tecnici e Professionali sono pressoché simili a quelle dei Licei, fatta eccezione per alcuni argomenti aggiuntivi. Si richiede,

perciò, una consapevolezza del pensiero matematico oltre che la capacità di padroneggiare il linguaggio formale e i concetti studiati, con la differenza che si punta maggiormente l'attenzione sugli aspetti applicativi della Matematica piuttosto che quelli dimostrativi.

Per il settore **Dati e previsioni**, il programma previsto dalle Linee guida risulta alquanto sostanzioso a fronte del quadro orario disponibile. Infatti se si va a vedere nel dettaglio, gli obiettivi formativi richiesti per questo nucleo di apprendimento sono numerosi.

- Richiami e integrazione di statistica descrittiva: dati statistici, grafici, i più importanti indici di posizione centrale (media aritmetica, media ponderata, moda e mediana) e di variabilità (campo variazione, varianza e scarto quadratico medio); calcolo degli indici (compresa la mediana) per dati raggruppati in classi; percentili, quartili e lo scarto interquartile; Studio congiunto di due caratteri: distribuzione di frequenza doppia, indipendenza statistica di due variabili; la connessione (dipendenza) tra mutabili statistiche e l'indice di Chi-Quadro di Pearson.

- L'interpolazione statistica e il metodo dei minimi quadrati; regressione e correlazione.

- Elementi di calcolo combinatorio: concetto di disposizioni, permutazioni e combinazioni semplici; la funzione fattoriale e i coefficienti binomiali.

- Probabilità: gli eventi; le concezioni classica e statistica della probabilità; l'impostazione assiomatica della probabilità; la probabilità della somma logica di eventi; eventi incompatibili; la probabilità condizionata; eventi stocasticamente indipendenti; la probabilità del prodotto logico degli eventi; il problema delle prove ripetute; il Teorema di Bayes.

- Concetto di variabile casuale (v.c.) discreta, di distribuzione di probabilità e di funzione di ripartizione; i valori caratterizzanti una v.c.

discreta; cenno alla distribuzione binomiale; le v.c. standardizzate; la v.c. continua e relativa funzione densità di probabilità, funzione di ripartizione, valor medio, varianza e deviazione standard.

- La distribuzione normale o gaussiana: la curva di densità e le sue caratteristiche (tra cui le sue misure di sintesi e le proprietà delle aree); la normale standardizzata e calcolo di probabilità con l'uso della tavola di Sheppard.

- Concetto di inferenza statistica; la popolazione e il campione e i rispettivi parametri; distribuzione della media campionaria e il Teorema del limite centrale; Concetto di intervallo di confidenza e stima per intervallo della media campionaria di grandi campioni.

Le abilità proposte dalle Linee guida sono anch'esse molteplici. Se si va ad esaminare nel dettaglio, si richiede la capacità di analizzare, classificare e rappresentare graficamente distribuzioni singole e doppie di frequenze, calcolare indici di posizione e di variabilità (sia con dati semplici che con dati raggruppati) nonché valutare l'indipendenza tra due caratteri, misurando il grado di connessione tra due variabili. Proseguendo, è richiesta la capacità di conoscere e saper ricavare la funzione interpolante lineare (retta dei minimi quadrati), oltre che l'abilità nel valutare la relazione e la dipendnza lineare tra due variabili (regressione e correlazione).
Si richiede di conoscere il concetto di disposizioni, permutazioni e combinazioni semplici, la funzione $n!$ e il coefficiente binomiale con l'aggiunta di saper utilizzarli in semplici equazioni e disequazioni.
In merito al calcolo delle probabilità, lo studente deve apprendere la concezione classica e statistica della probabilità, la definizione assiomatica e le relative proprietà, oltre che la capacità di saper determinare la probabilità di un evento utilizzando somma e/o prodotto logico. Conoscere la differenza tra eventi incompatibili ed indipendenti e inoltre saper determinare la probabilità di un evento utilizzando il teorema delle prove ripetute e quello di Bayes.
Continuando con il lato applicativo del calcolo delle probabilità si richiede la conoscenza del significato di variabile aleatoria discreta e di distribuzione di probabilità, sapendo calcolare la funzione di ripartizione, la media, la

varianza e la deviazione standard. Le medesime abilità sono richieste per l'argomento inerente alle variabili casuali continue.
Relativamente alla distribuzione normale, è richiesta la conoscenza di tutte le sue caratteristiche, nonché la capacità di saper utilizzare le tavole per la risoluzione di problemi diretti e inversi.
Per quanto riguarda la parte inferenziale è fondamentale conoscere il significato di inferenza statistica, di popolazione e di campione (con i rispettivi parametri) ed i teoremi relativi alla distribuzione della media campionaria. Infine viene richiesta la conoscenza del significato di stima intervallare, sapendo determinare un intervallo di confidenza di livello fissato.

Ciò che si può notare è una grande disparità di argomenti tra gli Istituti Tecnici e Professionali e i Licei, per quanto riguarda il settore disciplinare relativo alla Statistica. Data la scarsità oraria nella prima tipologia di scuole, i docenti si vedono obbligati a fare delle scelte in termini di argomenti da eliminare dal programma, nonché il vincolo di affrontarli in maniera limitata. Nella seconda tipologia, invece, il problema sembra marginale in quanto il quadro orario è superiore e gli obiettivi formativi sono in numero inferiore. Infatti l'indicazione principale per i Licei è "*pochi concetti e metodi fondamentali, acquisiti in profondità*".

1.4 Punti di forza e aspetti problematici del riordino

Sicuramente uno degli aspetti più problematici sorti con il riordino è la drastica riduzione dei quadri orari dedicati alla Matematica e le sempre più gravi ristrettezze, di personale e di risorse, in cui si dibattono la maggior parte delle istituzioni scolastiche. Inoltre l'aumento dei contenuti formativi, a fronte di un orario non adeguato, porta ad incentivare pratiche didattiche negative che gravano sulla formazione dello studente, puntando più sull'addestramento che sul reale apprendimento. Se si pensa poi al sovrannumero di studenti per classe (attorno ai 30 allievi) questo quadro orario risulta assolutamente

insufficiente per trattare in maniera adeguata tutti gli argomenti. Infatti nelle ore annue previste per l'insegnamento della Matematica devono essere svolte tutte le attività, comprese almeno due verifiche orali per quadrimestre. Un altro problema risiede nella traduzione didattica delle Indicazioni Nazionali e delle Linee guida, che può risultare non immediata. Pertanto molti insegnanti decidono di mantenere pressoché inalterati gli argomenti svolti, non ponendosi il problema delle novità introdotte dalla nuova riforma.
Invece se si vanno a guardare gli aspetti positivi del riordino della scuola, le nuove indicazioni nazionali/linee guida possiedono dei loro punti di forza che potrebbero essere individuati e sviluppati ai fini di un miglioramento sia a livello di apprendimento che di insegnamento della Matematica. Cogliere gli aspetti positivi significa offrire agli studenti e alle loro famiglie una migliore offerta formativa. Uno di questi aspetti potrebbe essere l'unità delle indicazioni per i vari tipi di liceo, garantendo un nucleo comune di conoscenze e competenze nella formazione matematica degli studenti che completeranno la scuola secondaria di II grado.
Un altro punto di forza (rilevante per questa tesi) riguarda l'inserimento in tutte le tipologie di scuola, con la stessa dignità dei temi più tradizionali, di argomenti di Statistica e Probabilità. Queste tematiche non erano presenti nei programmi dei Licei tradizionali e, invece, fanno parte degli obiettivi di apprendimento di tutti i sistemi scolastici europei. Inoltre è molto importante e significativa l'attenzione che viene data all'acquisizione del concetto di modello matematico che fornisce una ragione e un senso a tecniche che apparivano agli studenti come fini a se stesse.
Oltretutto, nelle indicazioni si invita più volte ad evitare inutili dispersioni in tecnicismi ripetitivi per evitare prassi didattiche inefficienti e inefficaci, al fine di conseguire un apprendimento soddisfacente. Uno degli scopi delle indicazioni nazionali è, appunto, quello di fornire una direzione al docente, in modo da ottimizzare il proprio lavoro in classe.

1.5 Gli insegnanti e le nuove indicazioni: testimonianza di una docente

A conferma di quanto detto in precedenza, parla una docente di Matematica di un istituto tecnico del vicentino, che insegna al triennio dell'indirizzo biochimico.

Il problema principale che l'insegnante riscontra nel riordino della scuola secondaria riguarda, proprio, il ridimensionamento dei quadri orari in parallelo con l'aumento dei contenuti formativi.

"*Con la riforma, le ore sono diminuite e le linee guida rilasciate dal Ministero hanno aumentato gli obiettivi formativi. Sono costretta, perciò, a fare tante cose e in poco tempo in maniera limitata e sintetica. Soprattutto per quanto riguarda Statistica, riesco solo a dare dei cenni ai miei studenti, quando invece mi piacerebbe approfondire determinati argomenti. Purtroppo ciò mi è impossibile.*"

La crescente importanza della Statistica nel mondo lavorativo, nonché universitario, porta all'esigenza di inserire tale disciplina nelle scuole, per dare una formazione più completa allo studente che si appresta a terminare il suo percorso di studio.

"*La Statistica è molto importante, soprattutto nal campo lavorativo. Sarebbe necessario avvicinare gli studenti a questa disciplina ma, soprattutto, sensibilizzare i miei colleghi all'importanza di questa scienza. Ad esempio, i miei ragazzi di quarta, tornati dallo stage, erano molto entusiasti perché durante il loro periodo di lavoro avevano visto più volte l'utilizzo di strumenti statistici, come la regressione. Nella scuola dove insegno, sono l'unica che decide di inserire un po' di Statistica nel programma. Gli argomenti sono tanti, talvolta eccessivi, e con le ore che ho disposizione mi limito solo a fare dei cenni. L'impegno più grosso è nelle classi quinte, dove il programma è molto sostanzioso; infatti di solito dedico tutto il secondo quadrimestre. Tuttavia,*

ribadisco, posso permettermi solo di fare cenni a determinati argomenti, soprattutto per quanto riguarda la distribuzione normale e l'inferenza. Il mio obiettivo in quinta è, appunto, arrivare a dare riferimenti sulla statistica inferenziale, che trovo sia molto utile in ambito lavorativo."

Tuttavia la maggior parte degli insegnanti di Matematica tende a concedere poca importanza e attenzione alla Statistica, preferendo eliminare argomenti inerenti a questa disciplina piuttosto che sacrificare concetti legati ad altri ambiti disciplinari, come ad esempio l'analisi matematica.

"*Sono d'accordo che analisi sia molto importante, specialmente per coloro che hanno intenzione di proseguire gli studi all'Università. Però sono dell'idea che come sia importante l'analisi matematica, lo sia anche la Statistica. Oltre a confrontarmi con i miei colleghi di matematica, spesso chiedo pareri anche ai miei colleghi di chimica organica e biochimica. Nelle loro materie si fa spesso uso di strumenti statistici, soprattutto regressione e inferenza. Perciò, proprio loro, mi chiedono di porre maggiore attenzione alla statistica nella mia materia, e ridurre il tempo che dedico agli argomenti nuovi previsti per analisi, come serie di funzioni ed equazioni differenziali. Mi piacerebbe che gli altri docenti di matematica vedessero l'importanza di questa materia e decidessero di dedicarle del tempo.*"

La necessità di sensibilizzare i docenti di Matematica all'insegnamento della Statistica nelle scuole ed avvicinare gli studenti a questa disciplina (per un futuro sia lavorativo che universitario) ha portato alla realizzazione del progetto PLS (Piano Lauree Scientifiche).
Lo scopo è proporre seminari per gli insegnanti, per indirizzarli nell'istruzione della Statistica e far comprendere l'importanza e la bellezza di questa scienza ai ragazzi delle scuole.

"*Credo che degli interventi tenuti da persone esterne all'istituzione scolastica, possano essere utili non solo ad avvicinare gli studenti ma anche a rendere consapevoli gli altri insegnanti, miei colleghi, dell'importanza che la*

Statistica ha nel mondo. A mio parere i contenuti di questi incontri dovrebbero essere intriganti ed interessanti per i ragazzi ai quali sono rivolti ma, cosa più importante, dovrebbero avere un legame con gli argomenti che gli studenti affrontano nel loro percorso di studio, in modo che capiscano l'utilità di quanto stanno studiando."

1.6 Lo scopo della tesi

La tesi propone due moduli didattici da proporre agli insegnanti delle scuole superiori, con un duplice scopo. Da una parte è importante avvicinare gli studenti alla Statistica in quanto molte sue tematiche si riscontrano in tutti i percorsi universitari di tipo scientifico; ma soprattutto è importante far comprendere loro l'importanza che tale disciplina gioca nell'affrontare problemi inerenti alla realtà che li circonda.
D'altro canto è necessario formare gli insegnanti suggerendo indicazioni su come affrontare argomenti di carattere statistico in classe, per fornire una preparazione di base agli studenti in linea con le proprie esigenze ed aspettative future.
Nel prossimo capitolo si cerca di dare una serie di linee guida e suggerimenti per orientare i docenti delle scuole superiori nell'insegnamento della Statistica in classe.

Capitolo 2

La Statistica in classe

2.1 Insegnare la Statistica a scuola

La scuola ha da sempre lo scopo di fornire agli studenti gli strumenti e le conoscenze per affrontare il loro futuro. Col passare degli anni il mondo che si trovano davanti è diventato sempre più complesso portando la necessità di possedere strumenti che permettano di prendere decisioni in condizioni di incertezza ed analizzare dati e informazioni di tipo quantitativo.
Per questa ragione la Statistica e il Calcolo delle Probabilità hanno assunto un ruolo sempre più importante nella formazione dei ragazzi, in quanto rappresentano gli strumenti più adeguati per prendere decisioni in situazioni di incertezza.
Come è stato precisato nel Capitolo 1, l'insegnamento della Matematica nella scuola secondaria di II grado ha subito un processo di revisione, nel quale si specifica che lo studente al termine dell'obbligo scolastico deve aver maturato una serie di competenze statistiche. Infatti si richiede allo studente la capacità di analizzare ed interpretare dati, sviluppando il ragionamento statistico, nonché saper raccoglierli, organizzarli e rappresentarli in maniera adeguata.
Risulta essenziale motivare gli studenti lavorando con i dati ed interagendo con le discipline dalle quali essi provengono, con lo scopo di "cucire l'abito mentale" di chi si appresta ad utilizzare la Statistica.

Quando si decide di proporre tematiche di carattere statistico in classe, un modo per indirizzare gli studenti in questa disciplina è rappresentare ed esplorare i dati mediante l'utilizzo strumentale della Matematica, ad esempio i modelli.
La Statistica è in grado di attivare diversi processi cognitivi dello studente:

- *logici*: trattazione formalizzata, propria della disciplina;
- *intuitivi*: trattazione legata all'esplorazione dei dati;
- *pianificatori ordinati*: schemi riassuntivi predisposti in maniera adeguata ed esercizi ripetuti;
- *emotivi*: discussione sui dati, collaborazione su progetti comuni e condivisione di esperienze. Questo tipo di attività è particolarmente gradito dai ragazzi e quindi adatto all'insegnamento della Statistica a scuola.

Il quesito che maggiormente preoccupa i docenti è:

Come far acquisire le competenze ai ragazzi?

La risposta a questa domanda va ricercata in ciò che attrae i ragazzi. Catturare la loro attenzione diventa più semplice affrontando temi che sono vicini alle loro esperienze quotidiane. Questo può incentivare la nascita di sane discussioni tra gli studenti, per sviluppare quel processo cognitivo citato precdentemente.
Gli strumenti informatici possono giocare un ruolo molto importante nell'apprendimento della Statistica. Sarebbe ideale che gli studenti imparino ad utilizzare un software statistico per grafici, tabelle, medie ed altri indici. Possedere le competenze per scegliere grafici e indici statistici appropriati, nonché saper descrivere l'evidenza presente nei dati risulta fondamentale, piuttosto che conoscere i dettagli computazionali. Per questo motivo l'approccio statistico alla conoscenza dei problemi legati a fenomeni reali si presta per lo più ad attività laboratoriali.

Il punto fragile del riordino risiede nella figura alla quale è affidato l'insegnamento della Statistica nella scuola, ossia il docente di Matematica. É fondamentale chiarire che il modo di pensare "statistico" risulta differente da quello "matematico" e questo rende le due discipline diverse tra loro.
La Statistica è una scienza strettamente legata ai dati, quindi non può prescindere dal contesto. Lo scopo principale è formare il docente di Matematica affinché sviluppi un modo di ragionare "statistico". Una formazione adeguata sarebbe in grado di dimostrare l'utilità della Statistica anche per l'insegnamento della Matematica ma, soprattutto, renderebbe consapevole il docente dell'importanza che tale disciplina gioca nell'avvicinare gli studenti e motivarli allo studio della Matematica.
Nonostante tutto, le due discipline sono strettamente legate in quanto la Statistica necessita degli strumenti matematici per raggiungere le proprie finalità. Quindi l'insegnamento collaborativo della Matematica e della Statistica contribuisce a motivare gli studenti allo studio di entrambe le discipline e, allo stesso tempo, fa apprezzare agli studenti il rigore della Matematica tanto quanto la potenza del metodo statistico per l'analisi quantitativa dei fenomeni reali.
Pertanto si può affermare che risulta necessaria l'armonizzazione fra le due discipline, rispetto agli elementi che le diversificano.
Il primo elemento di distinzione è legato al concetto di *dato*. Per la Matematica esistono solo numeri, mentre per la Statistica i dati sono numeri contestualizzati, frutto di un disegno di ricerca. Il *dato*, perciò, risulta fondamentale per la Statistica, ma non per la Matematica. In relazione a quanto detto si riscontra un'altra grande differenza tra le due discipline, ossia il concetto di *modello*. In un'ottica prettamente matematica, il modello si riduce ad una semplice formula, mentre in Statistica il suo concetto è più esteso. Infatti a fronte di obiettivi funzionali il modello, in senso statistico, è uno strumento che permette una riproduzione semplificata del fenomeno d'interesse, in modo da studiarne le caratteristiche e trarre conclusioni volte a rispondere a determinati quesiti. Molte sono le problematiche legate al modello, nella sua concezione statistica, come la scelta tra diversi modelli, per identificare quello migliore per l'analisi del fenomeno in questione, nonché il

problema di dover affrontare il contrasto con i dati.
L'insegnamento della Statistica deve essere veicolato in maniera adeguata da parte dei docenti, per garantire una formazione solida agli studenti. L'errore più comune è quello di spiegare concetti statistici attraverso un elenco di formule e grafici senza una contestualizzazione adatta. Inoltre il ragionamento statistico e i metodi quantitativi possono essere trasmessi anche attraverso altre discipline, come le scienze sperimentali e/o economico-sociali, per motivare gli studenti allo studio della materia dimostrandone l'utilità nell'affrontare tematiche legate alla realtà che li circonda. Le recenti teorie di apprendimento suggeriscono che una buona pratica di insegnamento consiste nel disegnare ambienti di apprendimento che stimolino i ragazzi a costruire la conoscenza.

2.2 Strumenti per la trasmissione delle competenze

Per indirizzare i docenti nell'insegnamento della Statistica in classe sono state proposte diverse esperienze, sia in ambito nazionale sia internazionale.
In generale un buon punto di partenza è consultare materiale didattico in rete, attraverso siti ufficiali dedicati alla divulgazione della cultura statistica. Ad esempio nel sito DiSIA dedicato all'insegnamento della Statistica e della Probabilità nella Scuola (`https://www.disia.unifi.it/`), vengono definite le competenze richieste agli studenti nella disciplina e si trova una sezione dedicata ad esempi concreti da utilizzare per introdurre la Statistica a scuola. Un altro sito utile è quello della SIS (Società Italiana di Statistica), una società scientifica senza fine di lucro che ha lo scopo fondamentale di promuovere lo sviluppo delle Scienze Statistiche e delle loro applicazioni in diversi ambiti disciplinari (`https://www.sis-statistica.it/`). Nel sito sono presenti diverse sezioni, tra le quali una dedicata alla didattica della Statistica, in cui è possibile scaricare materiali e informazioni utili per la divulgazione della disciplina. Sempre da questo sito è possibile raggiungere altre pagine web

interessanti volte a fornire ulteriori strumenti didattici e linee guida per i docenti. Molto importante è anche il sito dell'ISTAT (`http://www.istat.it`). A livello internazionale, l'*American Statistical Association*, è impegnata nell'aiutare i docenti di Matematica che devono insegnare concetti di Statistica. Il sito dell'associazione (`http://www.amstat.org/`) contiene progetti di lezione, corredati di dati e indicazioni pratiche, proposte e riviste dai docenti stessi. Un'altra organizzazione molto importante è l'associazione americana CAUSE, il cui scopo è quello di supportare e migliorare l'istruzione della Statistica attraverso diversi materiali come lezioni, dati, fumetti, canzoni e rompicapi.
I libri di testo scolastici contengono ancora molteplici errori, quindi è sconsigliata un'organizzazione didattica delle lezioni sulla base di questi. Per citarne alcuni:

- Baroncini, P., Manfredi, R. (2016). *MultiMath - Dati e previsioni.* Ghisetti e Corvi;
- Sasso, L. (2012). *La matematica a colori - edizione verde.* Petrini;
- Bergamini, M., Barozzi, G., Trifone, A. (2016). *Matematica.blu.* Zanichelli;

e molti altri.
È consigliabile al docente la preparazione autonoma del materiale facendo riferimento a siti specifici, nonchè a libri di testo extra-scolastici dedicati alla cultura statistica. Tra questi si consigliano:

- Agresti, A., Finlay, B. (2009). *Statistica per le Scienze Sociali.* Pearson;
- Agresti, A., Franklin, C. (2013). *Statistica: l'arte e la scienza d'imparare dai dati.* Pearson;
- Bernstein, S., Bernstein, R. (2003). *Statistica Descrittiva.* McGraw-Hill;
- Diamond, I., Jefferies, J. (2002). *Introduzione alla statistica. Per le scienze sociali.* McGraw-Hill;

- Rosenthal, J.S. (2005). *Le Regole del Caso: Istruzioni per l'Uso*. Longanesi;
- Spiegel (2003). *Probabilità e Statistica*. Schaum.

Nei prossimi capitoli vengono descritti due moduli didattici che possono essere utilizzati dai docenti di Matematica come esempi e linee guida per l'insegnamento della Statistica in classe.

Capitolo 3

Modulo didattico: correlazione e regressione

In questo capitolo viene descritto un primo modulo didattico da proporre ai docenti delle scuole superiori per l'insegnamento della Statistica in classe. Si tratta di moduli *self-consistent*, che gli insegnanti possono utilizzare per spiegare concetti statistici di base ai propri studenti. Lo scopo è indirizzare il docente nell'insegnamento della materia ma, allo stesso tempo, far vedere ai ragazzi l'utilità degli strumenti statistici tramite applicazioni a dati reali, che magari loro stessi hanno analizzato in altri ambiti disciplinari.
Il modulo proposto in questo capitolo è: *Correlazione tra Ammonio e Ferro nelle acque potabili del Veneto.*
Si tratta di un modulo didattico in cui si spiegano i concetti di correlazione e regressione semplice, applicandoli a un *dataset* (fonte: ARPAV) delle acque potabili della regione Veneto allo scopo di dimostare una correlazione positiva tra Ammonio e Ferro. In particolare, il *dataset* in questione è stato utilizzato dai ragazzi dell'indirizzo chimico-biologico dell'ITIS di Bassano del Grappa nell'anno scolastico 2017-2018, per l'analisi chimica degli elementi. In questo modo è possibile far apprezzare agli studenti l'interdisciplinarità della Statistica, utilizzando dati che loro stessi hanno analizzato in un altro contesto.

Il caso di studio

Il *dataset* è costituito dalle misurazioni relative al monitoraggio delle acque destinate al consumo umano in Veneto nell'anno 2011 (fonte: ARPAV). Le variabili del *dataset* rappresentano la quantità di quattro differenti componenti rilevate all'interno di campioni di acque potabili nel Veneto. Nello specifico, nel *dataset* si hanno le variabili: *As*, quantità di Arsenico; *Fe*, quantità di Ferro; *NH4*, quantità di Ammonio; *Mn*, quantità di Manganese, tutte misurate in mg/l.
Il *dataset* è costituito da *n=24* unità statistiche[1], che rappresentano 24 comuni della regione Veneto in cui sono state effettuate le misurazioni.
I dati sono organizzati secondo una tabella, nella quale le colonne sono rappresentate dalle 4 variabili e le righe dalle 24 unità statistiche. Ogni riga rappresenta la quantità delle componenti in ciascun comune del Veneto, mentre ogni colonna la quantità di ciascuna componente sui comuni considerati. Il *dataset* è:

	As	NH4	Fe	Mn
1	4.0	0.30	300	119
2	3.0	0.00	0	3
3	1.2	0.00	29	37
4	10.0	0.17	0	28
	...			
	...			
23	2.3	0.07	35	6
24	28.0	1.70	500	119

Scopo dell'analisi è valutare l'esistenza di una relazione tra Ferro e Ammonio nelle acque potabili del Veneto. Inoltre si intende valutare anche le relazioni tra altre componenti per capire se sono presenti altre relazioni rilevanti.

[1] Si definiscono *unità statistiche* le entità (individui, cose,...) che vengono osservate nello studio. L'insieme di tutte le unità statistiche di interesse per lo studio è detto *popolazione* di riferimento.

3.1 Concetti di base della Statistica: i dati e le variabili

I dati, come detto in precedenza, sono una raccolta di informazioni frutto di un disegno di ricerca e rappresentano uno dei fondamenti delle Scienze Statistiche. Senza i dati non si potrebbe fare Statistica.
Nelle analisi dei dati entra in gioco anche un altro concetto fondamentale, quello di *variabile*.
Una *variabile* descrive una caratteristica di interesse che viene rilevata sulle unità statistiche appartenenti al campione di riferimento, come esito di studio.
Ogni valore distinto che la variabile può assumere viene definito *modalità*.
Le variabili possono essere classificate in diverse tipologie. Una variabile, infatti, può essere:

- *qualitativa o categoriale*, quando le sue modalità sono espresse in forma verbale, ovvero le modalità sono dei valori non numerici (ad esempio: il genere o il credo religioso). A sua volta una variabile qualitativa può essere classificata come:
  - *sconnessa o nominale*, se tra le modalità non esiste nessun ordinamento (ad esempio: il tipo di scuola superiore o il colore degli occhi);
  - *ordinale*, se le modalità posseggono un ordine naturale, ovvero possono essere disposte lungo una scala (ad esempio: gli attributi "pessimo", "cattivo", "mediocre", "buono" e "ottimo" oppure i giorni della settimana).
- *quantitativa o numerica*, quando le modalità sono espresse da numeri (ad esempio: l'altezza o il numero di figli). A sua volta, una variabile quantitativa può essere classificata come:
  - *discreta*, quando l'insieme delle modalità è finito o numerabile, ovvero i suoi possibili valori possono essere elencati attraverso una

successione (ad esempio: il numero di figli, le pagine di un libro o i viaggi annuali);

– *continua*, quando l'insieme delle modalità è un intervallo, ovvero un sottoinsieme (eventualmente illimitato) dei numeri reali (ad esempio: il peso o l'altezza).

Relativamente al caso delle acque potabili nel Veneto, il *dataset* preso in considerazione è composto da *n=24* unità statistiche su cui sono state rilevate 4 variabili quantitative, che indicano le quantità di 4 elementi chimichi presenti nelle acque potabili del Veneto.

3.2 Analisi esplorativa

La prima fase di ogni analisi statistica consiste nell'organizzazione e la sintesi dei *dati*, costituiti dalle informazioni raccolte sulle unità statistiche che compongono il *campione*. A tale scopo si fa uso di vari strumenti, a partire da grafici e indici di sintesi.

3.2.1 Rappresentazione grafica dei dati: l'*istogramma*

In base al tipo di variabili si possono utilizzare diverse tipologie di grafico. Nel caso di variabili quantitative continue si utilizza l'*istogramma*.
Questo tipo di grafico è composto da rettangoli adiacenti le cui basi sono allineate su un asse orientato e dotato di unità di misura. L'adiacenza dei rettangoli vuole sottolineare una continuità del carattere. Ogni rettangolo ha base di lunghezza pari all'ampiezza della corrispondente classe mentre l'altezza è calcolata come densità di frequenza. Questa è data dal rapporto fra la frequenza (assoluta o relativa) associata alla classe e l'ampiezza della classe. Data una variabile, per *frequenza assoluta* di una modalità (o di un intervallo di valori) si intende il numero di unità statistiche che presentano quella modalità (o un valore della variabile nell'intervallo).

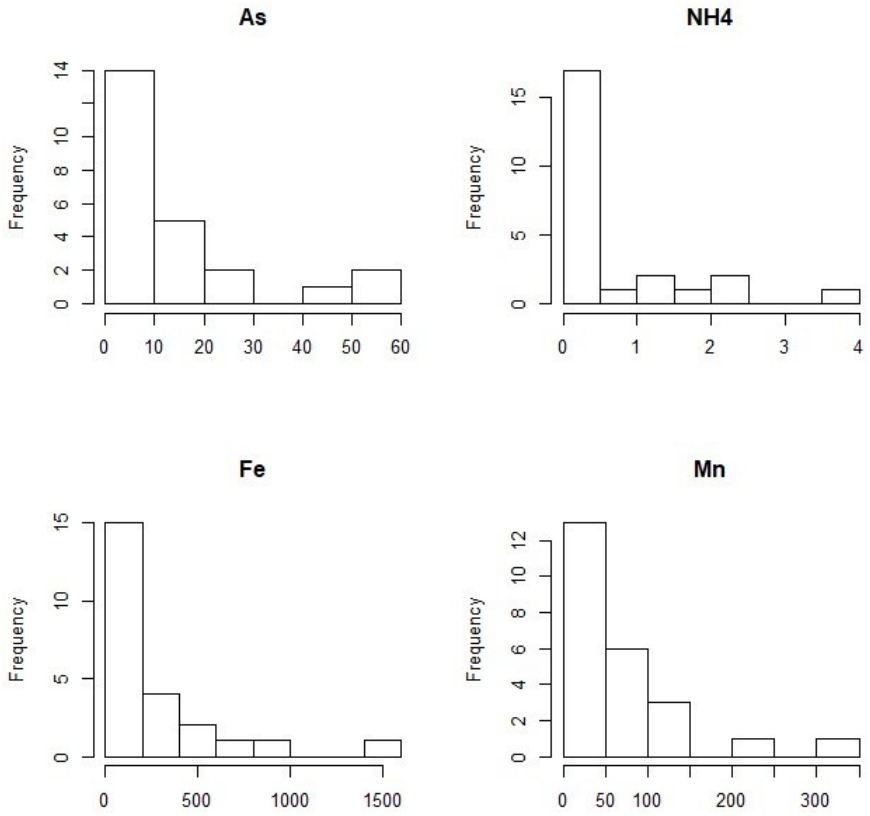


Figura 3.1: Istogrammi delle 4 variabili.

Dagli istogrammi in Figura 3.1 si può notare che le distribuzioni sono spostate principalmente verso sinistra, ossia verso valori piccoli delle variabili. Quindi le distribuzioni sono asimmetriche. Inoltre si osservano dei *valori anomali* (Figura 3.2). Nello specifico, tutti i rettangoli a destra rappresentano quei valori numericamente distanti dalla maggior parte delle osservazioni. Nelle analisi statistiche è importante evidenziare la presenza di valori anomali.

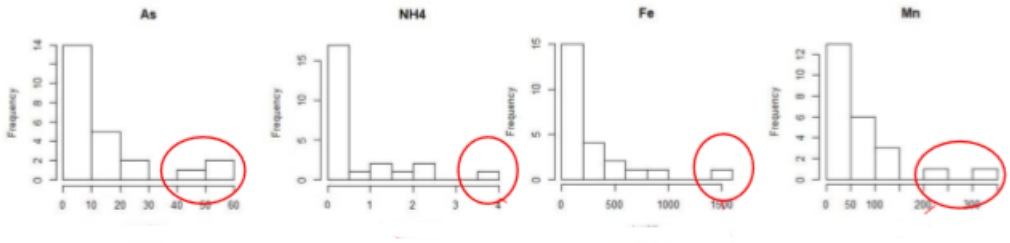


Figura 3.2: Osservazione dei valori anomali.

Oltre a rappresentazioni grafiche, nelle analisi per le variabili numeriche è utile avere a disposizione anche degli indici numerici di sintesi, che permettano di descrivere sinteticamente le caratteristiche delle osservazioni.

3.2.2 Indici di posizione

Si definisce *indice di posizione* una misura che descrive l'ordine di grandezza dei valori osservati, ossia il suo "centro". È quindi un singolo valore che si può ritenere "centrale" rispetto alla distribuzione di frequenza.
L'indice di posizione più comunemente utilizzato per variabili quantitative è la *media aritmetica*, che si calcola sommando i valori $(x_1, x_2, ..., x_n)$ osservati su n unità statistiche di una variabile quantitativa X di interesse e dividendo per il numero totale delle osservazioni:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{3.1}$$

dove $\sum$ è un simbolo matematico che abbrevia la somma di tutte le osservazioni x_i, per $i = 1, 2, ..., n$.

Proprietà della media aritmetica

- La media aritmetica è sempre compresa tra il più piccolo e il più grande dei valori osservati: $x_{min} \leq \bar{x} \leq x_{max}$, dove x_{min} è il valore minimo osservato e x_{max} il massimo;

- PROPRIETÀ DI BARICENTRO
  La somma degli scarti delle osservazioni dalla propria media $\bar{x}$ è zero[2]:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0;$$

- MINIMI QUADRATI
  La somma degli scarti al quadrato delle osservazioni da un valore c è minima se $c = \bar{x}$[3]:

$$\sum_{i=1}^{n}(x_i - c)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2;$$

- MEDIA DI TRASFORMAZIONI LINEARI
  La media di una trasformazione lineare dei dati è la stessa trasformazione lineare applicata alla media dei dati[4]:
  se $y_i = a + bx_i$, $i = 1, .., n \rightarrow \bar{y} = a + b\bar{x}$, con a e b numeri reali.

Prendendo in considerazione la variabile relativa al Ferro (Fe), la media aritmetica risulta:

$$\bar{x} = \frac{300 + 0 + 29 + 0 + ... + 0 + 23 + 35 + 500}{24} = 230.25\ mg/l.$$

Quindi si può affermare che nelle acque potabili del Veneto, in media sono presenti $230.25 mg/l$ di Ferro. Analogamente si procede con le altre variabili. I valori medi delle componenti risultano (in mg/l):

---

[2] *Dimostrazione*:
$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$

[3] *Dimostrazione*:
$\sum_{i=1}^{n}(x_i - c)^2 = \sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - c)]^2$
$= \sum_{i=1}^{n}[(x_i - \bar{x})^2 + (\bar{x} - c)^2 + 2(x_i - \bar{x})(\bar{x} - c)]$
$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2 + 2(\bar{x} - c)\sum_{i=1}^{n}(x_i - \bar{x})$
$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - c)^2.$
Infatti, per la proprietà di baricentro, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
La quantità $n(\bar{x} - c)^2$ è maggiore o uguale a zero (zero se $c = \bar{x}$).
Quindi, $\sum_{i=1}^{n}(x_i - c)^2$ è uguale a $\sum_{i=1}^{n}(x_i - \bar{x})^2$ più una quantità non negativa e questo dimostra la proprietà.

[4] *Dimostrazione*:
$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{b}{n}\sum_{i=1}^{n} x_i = \frac{na}{n} + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}.$

	Media
As	14.95
NH4	0.68
Fe	230.25
Mn	67.21

Si può osservare che il Ferro è la componente che in media risulta presente in maggiore quantità nelle acque venete, rispetto alle altre tre componenti. È opportuno osservare che la media aritmetica è molto sensibile alla presenza di valori anomali nei dati e questo rappresenta un suo difetto. Ad esempio, la media della variabile Fe è pari a 230.25 mg/l, ma togliendo i valori anomali diventa 175.04 mg/l. Per tenere conto della presenza di valori anomali può essere utilizzato, come indice di posizione, la *mediana*. La *mediana* di un insieme di dati ordinati in senso crescente è rappresentata dall'osservazione che occupa la posizione centrale. Per calcolare la mediana di n dati:

- si ordinano le n osservazioni in ordine crescente (o decrescente);
- se n è dispari la mediana corrisponde al valore centrale, ovvero all'osservazione che occupa la posizione
$$\frac{n+1}{2};$$
- se n è pari, la mediana è definita come la media dei due valori che occupano le posizioni
$$\frac{n}{2} \quad e \quad \frac{n+1}{2}.$$

Nel caso delle acque potabili del Veneto, la mediana relativa alla variabile Fe si calcola nel seguente modo:

1) ORDINAMENTO IN SENSO CRESCENTE DEI DATI

0 0 0 0 0 0 0 0 0 0 23 29 30 35 40 225 250 300 346 500 500 760 988 1500

2) CALCOLO DELLA MEDIANA

Il numero di osservazioni è pari (*n=24*) perciò la mediana viene stimata con la media delle osservazioni in posizione 12 e 13:

$$12^a osservazione = 29,$$

$$13^a osservazione = 30,$$

$$mediana = \frac{29 + 30}{2} = 29.5\, mg/l.$$

In maniera analoga si calcola la mediana delle altre variabili (in mg/l):

	Media	**Mediana**
As	14.95	10.00
NH4	0.68	0.28
Fe	230.25	29.50
Mn	67.21	32.50

Si può osservare che, per tutte le variabili, media e mediana hanno valori diversi tra loro, soprattutto per Ferro e Manganese. La mediana è preferibile alla media aritmetica quando si vuole evitare che eventuali dati anomali compromettano la stabilità dell'indice di posizione.
Esistono altri indici che generalizzano la mediana, ossia i *quantili*.
Il *quantile* di ordine α, con α numero reale nell'intervallo [0,1], è un valore Q_α che lascia alla sua sinistra almeno il $100\alpha\%$ delle osservazioni e alla sua destra il restante 100(1-α)%. I quantili con $\alpha = 0.25, 0.50, 0.75$ sono chiamati, rispettivamente, primo (Q_1), secondo (Q_2) e terzo (Q_3) *quartile* e dividono l'insieme dei dati in quattro parti uguali. Il secondo quartile coincide con la mediana. I quartili sono necessari per costruire un utile grafico: il *diagramma a scatola e baffi*.

Diagramma a scatola e baffi

Il *diagramma a scatola e baffi* (o *box-plot*) è una rappresentazione grafica utilizzata per descrivere la distribuzione di un campione di dati attraverso i

quantili.
Gli estremi della scatola sono definiti dal primo e il terzo quartile, la linea interna alla scatola è la mediana mentre i baffi, nella costruzione di base del diagramma, si estendono fino al valore minimo e massimo dei dati relativi. Esistono scelte alternative per rappresentare il diagramma, ma tutte

Figura 3.3: Diagramma a scatola e baffi.

utilizzano i tre quartili per rappresentare il rettangolo e differiscono per la lunghezza dei baffi. Una possibilità è prendere come estremi del digramma i valori $(Q_1\text{-}Q_3)/2$ e $(Q_3\text{-}Q_1)/2$ in modo che entrambi i baffi siano lunghi 3/2 volte la lunghezza della scatola. Tutti i valori che fuoriescono dai limiti vengono considerati valori anomali.
La Figura 3.4 riporta i box-plot delle 4 variabili del *dataset* relativo alle acque potabili del Veneto.

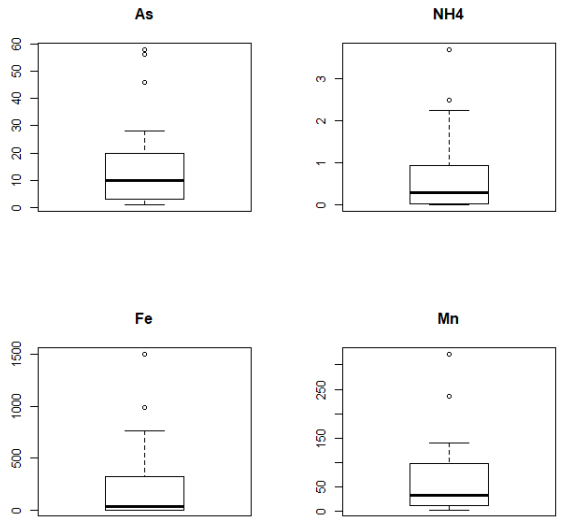


Figura 3.4: Box-plot di Arsenico, Ammonio, Ferro e Manganese.

Come osservato negli istogrammi, le distribuzioni delle variabili sono caratterizzate da un'asimmetria a sinistra e presentano dei valori anomali.

3.2.3 Indici di variabilità

Mediante gli indici di posizione si cerca di sintetizzare una distribuzione statistica, con lo scopo di raccogliere parte rilevante delle informazioni contenute in essa. Tuttavia un indice di posizione (qualunque esso sia) non è sufficiente a sintetizzare completamente la distribuzione, perchè si trascura un aspetto importante: la *variabilità*. La *variabilità* esprime la tendenza delle unità statistiche a manifestarsi con modalità diverse del carattere. Fondamentale è l'analisi di quest'attitudine soprattutto quando si confrontano distribuzioni diverse, in quanto, a parità di media, esse possono risultare diverse in termini di dispersione.
Esistono diversi indici che servono a misurare la variabilità di un carattere,

ma il più utilizzato per variabili quantitative è la *varianza*, data da

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.2}$$

Esiste una formula alternativa per il calcolo della varianza data dalla differenza della media dei quadrati delle osservazioni e il quadrato della media:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{3.3}$$

La varianza misura la distanza dei dati dalla media aritmetica, valutata attraverso i quadrati delle differenze.
Siccome l'unità di misura della varianza è data dal quadrato dell'unità di misura con cui sono stati rilevati i dati, per ottenere un indice di variabilità sulla scala dei dati originari si considera spesso la radice quadrata della varianza, che viene detta *deviazione standard* (o *scarto quadratico medio*):

$$\sigma_x = \sqrt{\sigma_x^2}. \tag{3.4}$$

Proprietà della varianza

- SEGNO DELLA VARIANZA
  La varianza non è mai negativa, ed è zero se e solo se tutte le unità presentano uguale modalità del carattere;

- VARIANZA DI TRASFORMAZIONI LINEARI[5]
  La varianza di una trasformazione lineare dei dati non dipende dall'intercetta ma solo dal coefficiente angolare della trasformazione:
  per $y_i = a + bx_i$, $i = 1, .., n \rightarrow \sigma_y^2 = b^2\sigma_x^2$, con a e b numeri reali.

[5] *Dimostrazione*:
Sia $y_i = a + bx_i$, $i = 1, ..., n$.
$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}[a + bx_i - (a + b\bar{x})]^2 = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\bar{x})^2$
$= \frac{1}{n}\sum_{i=1}^{n} b^2(x_i - \bar{x})^2 = b^2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = b^2\sigma_x^2$.

Si procede con il calcolo della varianza e della deviazione standard per la variabile *Fe*.

1) VALORI DELLA VARIABILE Fe

```
300 0 29 0 500 0 0 0 250 0 30 988 225 0 40 760 1500 0 0 346 0 23
35 500
```

2) VALORI AL QUADRATO DELLA VARIABILE Fe

```
90000 0 841 0 250000 0 0 0 62500 0 900 976144 50625 0 1600 577600
2250000 0 0 119716 0 529 1225 250000
```

3) CALCOLO VARIANZA

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{90000 + 0 + 841 + 0 + 250000 + ... + 1225 + 250000}{24} = 192986.7$$

$$\bar{x} = 230.25\, mg/l$$

$$\sigma_x^2 = 192986.7 - 230.25^2 = 139971.6\, mg^2/l^2$$

$$\sigma_x = \sqrt{139971.6} = 374.1278\, mg/l.$$

Un altro indice di variabilità è lo *scarto interquartile*, calcolato sottraendo dal terzo quartile il primo quartile e comprende, dunque, il 50% delle osservazioni:

$$Scarto\ interquartile = Q_3 - Q_1. \tag{3.5}$$

Tale indice risulta più robusto in presenza di valori anomali e nel box-plot corrisponde all'ampiezza della scatola.
Di seguito vengono riportati gli indici di posizione e variabilità delle 4 variabili considerate.

	Media	**Mediana**	**Varianza**	**Dev. Std.**	**Scarto Inter.**
As	14.95	10.00	271.17	16.47	17
NH4	0.68	0.28	0.91	0.96	0.73
Fe	230.25	29.50	139971.60	374.13	311.5
Mn	67.21	32.50	5950.91	77.14	86

In questo caso si può osservare che le variabili relative al Ferro e al Manganese hanno una variabilità maggiore, rispetto alle altre.

3.3 Analisi bivariata dei dati

In Statistica, in molte situazioni, si è interessati a studiare se esiste una relazione tra due variabili misurate sulle stesse unità. Ad esempio in questo caso di studio ci si chiede:

Le misurazioni relative al Ferro (Fe) sono in relazione
con le misurazioni relative all'Ammonio (NH4)?

Nello specifico si è interessati a capire se esiste una relazione positiva tra le due componenti.
Un altro obiettivo che ci si può porre è prevedere il valore di una variabile conoscendo il valore di un'altra.
La Statistica fornisce diversi strumenti per rispondere a questo tipo di domande, adatti alla natura delle variabili in esame. Per le variabili quantitative si hanno a disposizione:

- la **CORRELAZIONE**, che misura la dipendenza lineare tra due variabili;
- la **REGRESSIONE**, che valuta un modello lineare tra due variabili.

3.3.1 La correlazione

Per *correlazione* si intende la misura dell'associazione tra due variabili statistiche, che si utilizza quando si hanno a disposizione coppie di valori di variabili. Non si tratta necessariamente di un rapporto di causa-effetto, ma semplicemente della tendenza di una variabile a variare in funzione di un'altra.

Gli strumenti utili per valutare la relazione tra due variabili sono:

- il *diagramma di dispersione*, che fornisce una valutazione grafica;
- il *coefficiente di correlazione*, che quantifica il grado di correlazione (ossia il segno e la forza di una relazione lineare).

Il *diagramma di dispersione* è un grafico in cui i valori osservati su due variabili sono riportati su un piano cartesiano. Il grafico di dispersione può essere utile per visualizzare il grado di correlazione, cioè di dipendenza lineare.
Nell'esempio del monitoraggio delle acque potabili nel Veneto, si può esprime la quantità di Ammonio (Y) in funzione della quantità di Ferro (X), misurate in mg/l (Figura 3.5). Dal grafico di dispersione si osserva che all'aumentare dell'una aumenta anche l'altra, ad indicare una *correlazione positiva*.

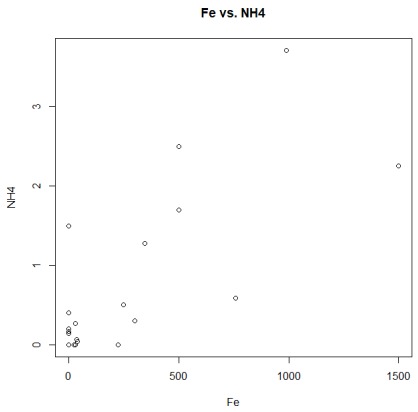


Figura 3.5: Diagramma di dispersione Fe vs. NH4.

Lo studio della correlazione è di notevole interesse perché permette di individuare relazioni tra variabili.
Quando si parla di correlazione bisogna prendere in considerazione due aspetti: il *tipo di relazione* esistente tra due variabili e la *forma* della relazione.

Per quanto riguarda il *tipo di relazione*, si può distinguere tra relazione *lineare* e *non lineare*. Per comprendere che tipo di relazione lega i dati è necessario rappresentarli in un diagramma di dispersione:

- la relazione è di tipo lineare se l'andamento dei dati può essere descritto sotto forma di una retta;
- la relazione è di tipo non lineare se si riscontra un andamento curvilineo dei dati (es. parabola o iperbole).

In merito alla *forma* della relazione è necessario fare una distinzione tra *forza* e *direzione*.
La *direzione* può essere di due tipologie:

- *positiva*, se all'aumentare dei valori di una variabile si osserva un aumento dei valori dell'altra;
- *negativa*, se all'aumentare dei valori di una variabile si osserva una diminuzione dei valori dell'altra.

La *forza* si riferisce all'entità della relazione esistente tra due variabili.
Quanto più i punti sono allineati attorno ad una retta, tanto più forte è la relazione lineare tra le due variabili. Al contrario, se i dati sono dispersi in maniera uniforme, significa che non esiste alcuna relazione lineare tra le due variabili.

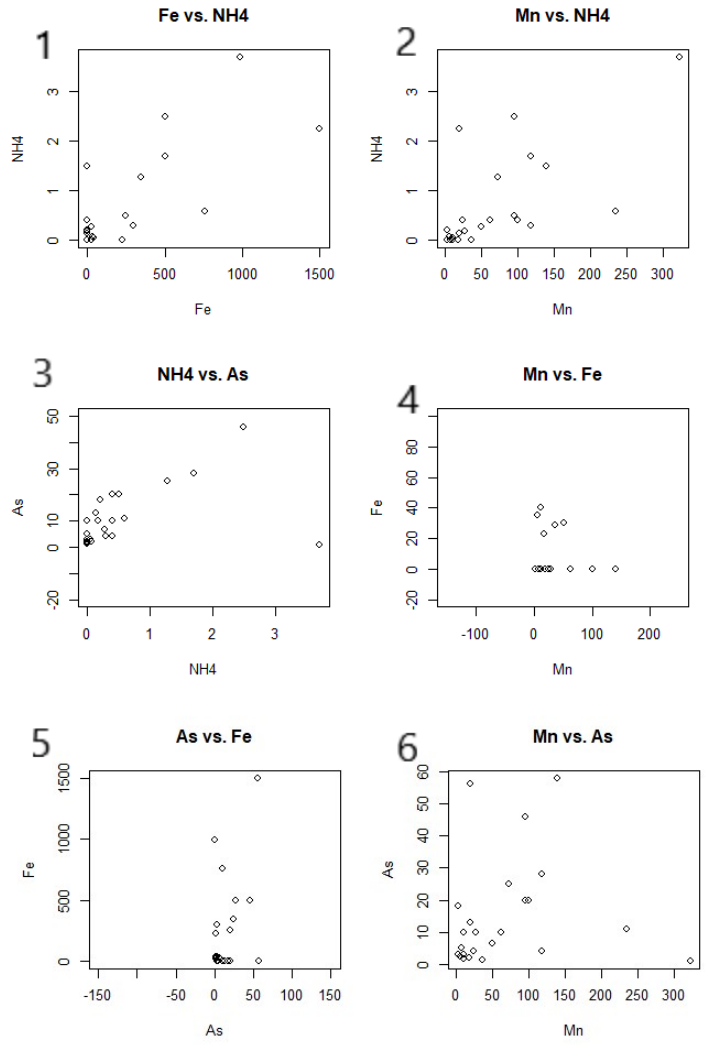


Figura 3.6: Diagrammi di dispersione tra le variabili del *dataset*.

I diagrammi di dispersione in Figura 3.6 sono essenziali per analizzare la correlazione tra le componenti delle acque potabili nel Veneto, in particolare quella tra Ferro e Ammonio.
Dai grafici 1-2 si può osservare che all'aumentare della variabile X (asse orizzontale) aumenta anche la variabile Y (asse verticale). Questo sta ad indicare che sussiste una relazione positiva tra le due variabili, seppure i punti tendano a disperdersi nel piano. Sarà necessario quantificare la forza della relazione attraverso l'indice di correlazione.

Dai grafici 4-5 si può osservare che non emerge una relazione lineare tra le due variabili. Infatti non si osserva un andamento sistematico dei punti, che possa essere espresso mediante una funzione matematica.
Dal grafico 6 appare una debole relazione positiva tra le due variabili, in quanto la nuvola di punti ha un'elevata dispersione.
Infine dal grafico 3 si può riscontrare un lieve andamento positivo tra le due variabili, anche se la nuvola di punti risulta sostanzialmente piatta.

Per quantificare il grado di associazione tra due variabili quantitative, viene utilizzato un indice che misura la dispersione nel piano dei punti dal proprio centro, $(\bar{x}, \bar{y})$: la *covarianza*.
La covarianza si basa sugli scarti dei valori osservati di X e Y, rispettivamente, dalle proprie medie:

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{3.6}$$

La covarianza misura la direzione della relazione lineare tra le due variabili, in modo da capire se si muovono entrambe verso la stessa direzione o in direzioni opposte. Il segno della covarianza risulta coerente con il senso crescente o decrescente dell'andamento dei dati. Infatti assume valore positivo in caso di concordanza, ossia X e Y crescono o decrescono insieme; altrimenti assume valore negativo in caso di discordanza, ossia quando una delle due variabili cresce mentre l'altra decresce e viceversa (Figura 3.7).

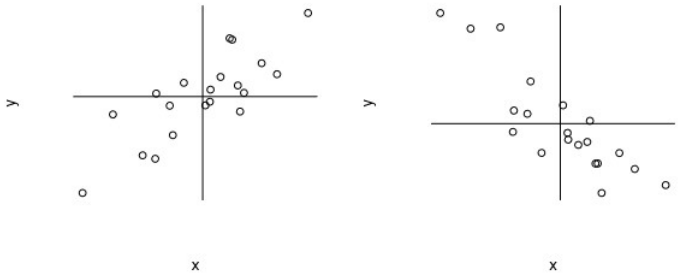


Figura 3.7: Grafico esplicativo della covarianza.

Proprietà della covarianza

1. SIMMETRIA
   $\sigma_{xy} = \sigma_{yx}$;

2. COVARIANZA DI TRASFORMAZIONI LINEARI[6]
   se $x_i^* = a + bx_i$, $i = 1, .., n \rightarrow \sigma_{x^*y} = b\sigma_{xy}$;

3. COVARIANZA DI UNA VARIABILE CON SÈ STESSA[7]
   $\sigma_{xx} = \sigma_x^2$;

4. DOMINIO: $-\sigma_x\sigma_y < \sigma_{xy} < \sigma_x\sigma_y$[8].

Esiste una formula semplificata per il calcolo della covarianza data da:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Per esprimere la relazione esistente tra due variabili numeriche, in termini di entità e direzione, si utilizza il **coefficiente di correlazione**. Questo indice, con dominio limitato, è molto utile perchè permette di quantificare la forza della relazione.

---

[6] *Dimostrazione*:
$\sigma_{x^*y} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(bx_i - b\bar{x})(y_i - \bar{y})$
$\frac{1}{n}\sum_{i=1}^{n} b(x_i - \bar{x})(y_i - \bar{y}) = b\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = b\sigma_{xy}$.

[7] *Dimostrazione*:
$\sigma_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sigma_x^2$.

[8] *Dimostrazione*:
Indichiamo con $var(X) = \sigma_x^2$. Calcoliamo la seguente varianza: $var(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y})$.
$var(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}) = var(\frac{X}{\sigma_x}) + var(\frac{Y}{\sigma_y}) + 2cov(\frac{X}{\sigma_x}, \frac{Y}{\sigma_y})$
$\frac{1}{\sigma_x^2}var(X) + \frac{1}{\sigma_y^2}var(Y) + 2\frac{1}{n}\sum_{i=1}^{n}[(\frac{x_i}{\sigma_x} - (\frac{\bar{X}}{\sigma_x}))(\frac{y_i}{\sigma_y} - (\frac{\bar{Y}}{\sigma_y}))]$
$= \frac{\sigma_x^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_y^2} + 2\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_x\sigma_y}(x_i - \bar{x})(y_i - \bar{y})$
$= 2 + 2\frac{1}{\sigma_x\sigma_y}[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})] = 2(1 + \frac{\sigma_{xy}}{\sigma_x\sigma_y})$.
Poichè quest'ultima quantità ottenuta non può essere negativa, si ha:
$2(1 + \frac{\sigma_{xy}}{\sigma_x\sigma_y}) \geq 0$
$\frac{\sigma_{xy}}{\sigma_x\sigma_y} \geq -1 \rightarrow \sigma_{xy} \geq -\sigma_x\sigma_y$.
Procedendo in modo analogo, calcolando $var(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y})$ si ottiene che $\sigma_{xy} \leq \sigma_x\sigma_y$.

Il coefficiente di correlazione è definito come

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{j=1}^{n}(y_j - \bar{y})^2}}, \tag{3.7}$$

$$-1 \leq \rho_{xy} \leq 1,$$

con σ_{xy} covarianza tra le variabili X e Y e σ_x e σ_y le rispettive deviazioni standard.

Tale coefficiente ammette valori nell'intervallo $[-1, 1]$ a seconda della forza della relazione. Assume valore -1 quando si presenta una perfetta correlazione negativa, mentre è pari a $+1$ quando si riscontra una perfetta correlazione positiva. Una correlazione uguale a 0 indica che tra le due variabili non sussiste alcun tipo di relazione lineare. In sintesi:

- $\rho_{xy} = 1 \rightarrow$ correlazione positiva perfetta (tutti i punti stanno su una retta: concordi);
- $\rho_{xy} =$ -1 $\rightarrow$ correlazione negativa perfetta (tutti i punti su una retta: discordi);
- $\rho_{xy} > 0 \rightarrow$ correlazione positiva;
- $\rho_{xy} < 0 \rightarrow$ correlazione negativa;
- $\rho_{xy} = 0 \rightarrow$ assenza di correlazione lineare.

Il coefficiente ρ_{xy} misura il grado di correlazione tra le variabili solo se questa è di tipo lineare. Un valore pari a 0 di tale coefficiente indica un'assenza di relazione lineare ma non comporta un'assenza di legame tra le due variabili. Infatti la relazione potrebbe essere di un'altra tipologia (es. curvilinea come in Figura 3.8).

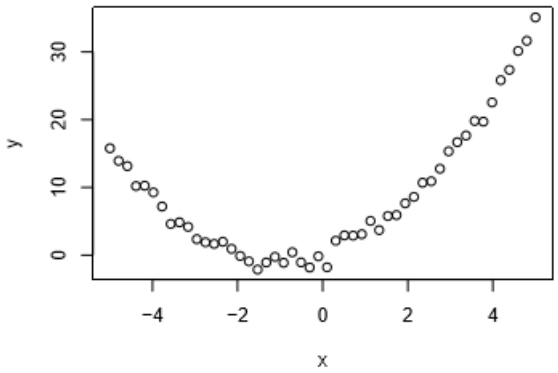


Figura 3.8: Esempio di relazione curvilinea.

Lo scopo dell'analisi è misurare la correlazione tra Ferro e Ammonio all'interno delle acque potabili. Di seguito vengono riportati i passaggi per il calcolo del coefficiente di correlazione tra le variabili *Fe* e *NH4*, con $X = Fe$ e $Y = NH4$.
Gli ingredienti per il calcolo del coefficiente di correlazione sono:
$\bar{x} = 230.25$
$\bar{y} = 0.6841667$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 3359319$
$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 21.95698$
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 6478.205$
Il coefficiente di correlazione risulta:

$$\rho_{xy} = \frac{6478.205}{\sqrt{3359319 * 21.95698}} = 0.75.$$

Tale valore indica una correlazione positiva tra Ferro e Ammonio, come ci si aspettava dal grafico di dispersione e come ipotizzato dai docenti di Chimica. Oltre al legame tra Ferro e Ammonio risulta interessante andare ad osservare anche la relazione tra le altre componenti. La Tabella 3.1 riporta le correlazioni tra le 4 variabili.

	As	**NH4**	**Fe**	**Mn**
As	1.00	0.54	0.43	0.13
NH4	0.54	1.00	0.75	0.68
Fe	0.43	0.75	1.00	0.51
Mn	0.13	0.68	0.51	1.00

Tabella 3.1: Tabella con le correlazioni.

Dalla Tabella 3.1 si osserva un'elevata correlazione positiva tra Ammonio e Manganese. Una correlazione debole si registra per Arsenico e Manganese, mentre le altre correlazioni risultano positive seppure di intensità minore rispetto a Fe-NH4 e Mn-NH4.

3.3.2 La regressione

Lo scopo della regressione è stimare un'eventuale relazione funzionale esistente tra la variabile risposta Y e la variabile esplicativa X. Quindi si vuole capire come la variabile X influenzi la variabile Y, oppure prevedere la risposta a partire dai valori della variabile esplicativa.

Il modello di regressione lineare semplice

Il modello di regressione lineare semplice è il modello più semplice per analizzare la relazione tra una variabile dipendente e una variabile indipendente. Nell'equazione di regressione la variabile dipendente è una funzione della variabile indipendente, più un termine d'errore che rappresenta ciò che il modello non spiega. In particolare, un modello di regressione lineare assume la seguente forma:

$$Y = \alpha + \beta X + \epsilon, \tag{3.8}$$

dove α rappresenta l'intercetta della retta di regressione, β il coefficiente angolare e ϵ il termine d'errore.
Si pone quindi il problema di come determinare un valore ragionevole per i parametri α e β. Il metodo più comunemente utilizzato per ottenere la stima del modello è il metodo dei *minimi quadrati*. Tale metodo prevede di

determinare i parametri α e β in modo che la retta sia il più vicino possibile alle osservazioni. Tale retta è detta *retta dei minimi quadrati*, ossia quella retta che rende minimi gli scarti al quadrato dai dati osservati.
Vengono definiti i *valori previsti* come:

$$\tilde{y_i} = \alpha + \beta x_i,$$

ossia quei valori che la variabile Y dovrebbe assumere a partire da $X = x_i$ nel caso in cui la relazione tra le due variabili sia quella lineare, $i = 1, ..., n$.
Quindi per stimare i parametri si minimizza la somma delle distanze tra i valori osservati e i valori previsti, detta *somma dei minimi quadrati*:

$$s^2(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \tilde{y_i})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{3.9}$$

Nella quantità $s^2(\alpha, \beta)$ le incognite da determinare sono α e β, mentre i valori x_i e y_i sono valori osservati.
La retta costituita dai valori stimati di α e β che minimizzano $s^2(\alpha, \beta)$ viene detta *retta dei minimi quadrati*. I valori che rendono minima la somma dei minimi quadrati risultano[9]:

$$\hat{\beta} = \frac{\sigma_{xy}}{\sigma_x}, \tag{3.10}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}. \tag{3.11}$$

[9] *Dimostrazione*:
Posto $y_i^* = y_i - \beta x_i, i = 1, ..., n$, la somma dei minimi quadrati $s^2(\alpha, \beta)$ può essere riscritta come $\sum_{i=1}^{n} (y_i^* - \alpha)^2$. Sfruttando la proprietà dei minimi quadrati della media aritmetica la quantità $\sum_{i=1}^{n} (y_i^* - \alpha)^2$ è minima per:
$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i^* = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i) = \frac{1}{n} y_i - \beta \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{y} - \beta\bar{x}$.

Sostituendo il valore ottenuto in $s^2(\alpha, \beta)$ si ottiene:
$\sum_{i=1}^{n} (y_i - \bar{y} - \beta x_i + \beta\bar{x})^2 = \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta(x_i - \bar{x})]^2 =$
$= \sum_{i=1}^{n} (y_i - \bar{y})^2 + \beta^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\beta \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) =$
$= n\beta^2\sigma_x^2 - 2n\beta\sigma_{xy} + n\sigma_y^2.$

Si tratta di una funzione quadratica in β, il cui grafico è una parabola con concavità rivolta verso l'alto. Tale funzione è minima nel suo vertice, ossia per

$$\hat{\beta} = \frac{-(-2n\sigma_{xy})}{2n\sigma_x^2} = \frac{\sigma_{xy}}{\sigma_x^2}.$$

Nel caso di studio relativo a Ferro e Ammonio la (3.6) diventa:

$$NH4 = \alpha + \beta Fe + \epsilon. \tag{3.12}$$

I valori stimati per i parametri risultano:

$$\hat{\beta} = \frac{6478.205}{3359319} = 0.0019,$$

$$\hat{\alpha} = 0.6841667 - 0.0019 * 230.25 = 0.24.$$

Quindi la retta di regressione stimata è data da:

$$\hat{NH4} = 0.24 + 0.0019 \times Fe.$$

Come si interpretano i valori ottenuti?

- $\hat{\alpha}$ rappresenta la media della variabile risposta Y, assumendo che la variabile esplicativa X sia nulla;
- $\hat{\beta}$ rappresenta la variazione in media della variabile Y al variare unitario della variabile X.

Quindi si può affermare che il livello medio di Ammonio rilevato all'interno delle acque potabili è pari a circa 0.24 mg/l quando non c'è presenza di Ferro. Mentre si osserva un aumento dello 0.19% dell'Ammonio all'aumentare unitario del Ferro.
La retta riportata in Figura 3.10 è utile anche per fare previsioni sulla variabile risposta. Ad esempio, per $X = 250$ si trova $\hat{Y} = 0.24+0.0019\times250 = 0.715$. Quindi in corrispondenza di 250 mg/l di Ferro si prevedono 0.715 mg/l di Ammonio.

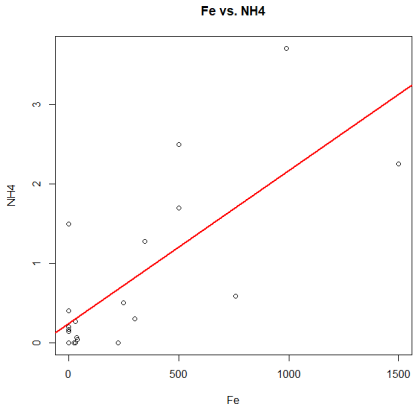


Figura 3.9: Retta di regressione stimata tra Fe e $NH4$.

Come si valuta se la retta si adatta bene ai dati?

Esiste un indice che sintetizza l'adattamento generale della retta di regressione ai dati. Tale indice si chiama *Coefficiente di determinazione*[10].

[10]Interpretazione di R^2 come proporzione di varianza spiegata:

- Siano $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i, i = 1, .., n$, i valori stimati con la retta dei minimi quadrati.
- Si ha:
  $\sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i) =$
  $\sum_{i=1}^n (y_i - \bar{y}) - \hat{\beta}\sum_{i=1}^n (x_i - \bar{x}) = 0.$
- Inoltre si ha:
  $\sum_{i=1}^n (y_i - \hat{y}_i)x_i = \sum_{i=1}^n (y_i - \hat{y}_i)(x_i - \bar{x}) = \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i)(x_i - \bar{x}) =$
  $n\sigma_{xy} - \hat{\beta}n\sigma_x^2 = 0.$
- Dall'identità $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i + \hat{y}_i - \hat{y}_i - \bar{y}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
  $+2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$, utilizzando le due relazioni precedenti, si vede che l'ultima sommatoria è nulla. Quindi $\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ cioè,

  $$VARIANZA_{TOTALE} = VARIANZA_{RESIDUA} + VARIANZA_{SPIEGATA}$$

- Si vede che $R^2 = \frac{VARIANZA_{SPIEGATA}}{VARIANZA_{TOTALE}}$.
  Infatti $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\bar{y} - \hat{\beta}\bar{x} + \hat{\beta}x_i - \bar{y})^2 = \frac{n\hat{\beta}^2\sigma_{xy}^2}{\sigma_x^2}$, perciò si ha:
  $$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{n\sigma_{xy}^2}{n\sigma_x^2\sigma_y^2} = R^2.$$

Per calcolare tale indice, indicato con R^2, è sufficiente far riferimento al coefficiente di correlazione ρ_{xy} e, siccome non ha importanza la direzione della correlazione (positiva o negativa), si può elevare ρ_{xy} al quadrato:

$$R^2 = \rho_{xy}^2. \tag{3.13}$$

Il coefficiente di determinazione assume valori nell'intervallo $[0, 1]$ a seconda della bontà di adattamento del modello ai dati:

− se $R^2 = 1 \rightarrow$ adattamento perfetto, ossia tutti i punti stanno sulla retta;
− se $R^2 = 0 \rightarrow$ il modello non si adatta per niente ai dati;
− come regola empirica se $R^2 = 0.8 \rightarrow$ il livello di adattamento è "buono".

In merito alla relazione tra Ferro e Ammonio, il coefficiente di determinazione del modello stimato risulta $R^2 = 0.754^2 = 0.57$. Un valore sostanzialmente basso di tale coefficiente, segnale della molta dispersione attorno alla retta.

Un'altra relazione interessante da approfondire è quella esistente tra Ammonio e Manganese. La retta di regressione stimata risulta:

$$\hat{NH4} = 0.119 + 0.008 \times Mn. \tag{3.14}$$

Dal modello stimato si può affermare che il livello medio di Ammonio rilevato all'interno delle acque potabili è pari a circa 0.119 mg/l quando non si riscontra presenza di Manganese. Per di più si osserva un aumento del 0.8% dell'Ammonio all'aumentare unitario del Manganese.
In Figura 3.9 viene riportato la retta di regressione stimata. Il grafico non si discosta molto dalla precedente. La relazione tra Ammonio e Manganese risulta positiva anche se la retta sembra non cogliere sufficientemente bene l'andamento dei dati. Infatti $R^2 = 0.68^2 = 0.46$. Il modello si adatta in maniera peggiore rispetto al precedente. Infatti dal grafico si osserva che i punti si disperdono molto intorno alla retta.

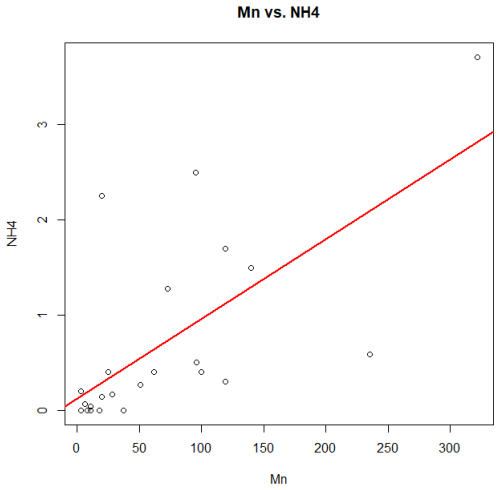


Figura 3.10: Retta di regressione stimata tra Ammonio e Manganese.

Sono state condotte anche le analisi relative alle relazioni tra le altre componenti rilevate. In particolare sono state prese in considerazione le seguenti relazioni:

- As vs. NH4
- Mn vs. Fe
- As vs. Fe
- Mn vs. As

Le rette stimate sono:
$\hat{NH4} = 0.21026 + 0.03170 \times As$
$\hat{Fe} = 63.9289 + 2.4747 \times Mn$
$\hat{Fe} = 84.506 + 9.749 \times As$
$\hat{As} = 13.10951 + 0.02738 \times Mn$

I grafici relativi alle rette stimate sono presentati in Figura 3.11.

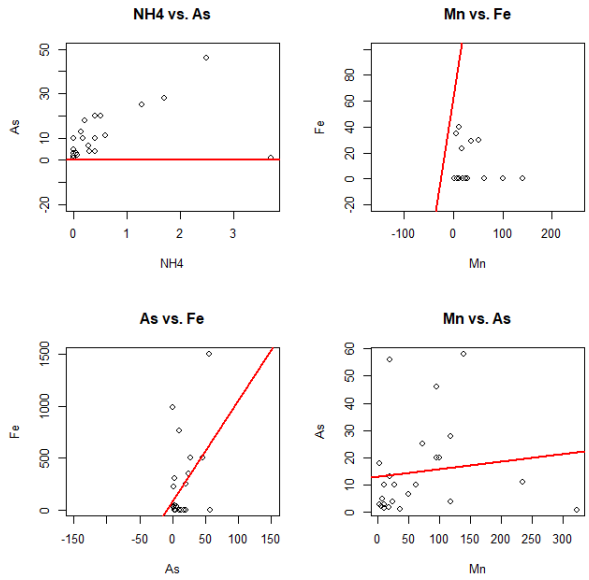


Figura 3.11: Rette di regressione stimate.

Dal primo grafico (in alto a sinistra) si può osservare che la retta è praticamente piatta anche se $\rho_{xy} = 0.54$. Il grafico indica l'assenza di relazione tra i due elementi, tuttavia un tale valore di ρ_{xy} suggerisce la presenza di una relazione lineare (probabilmente sono presenti dei *punti influenti*[11]). Anche la retta del quarto grafico (in basso a destra) tende ad essere piatta, seppur leggermente inclinata positivamente. Si può affermare che sussiste una debole relazione positiva ($\rho_{xy} = 0.13$). Anche per Manganese e Arsenico si può concludere che non si riscontra una relazione rilevante. Rispetto al caso precedente il coefficiente di correlazione risulta molto più basso, sebbene entrambi i grafici non mostrino evidenza di una relazione lineare. Probabilmente nel primo caso ci sono dei *punti influenti* che, se individuati, portano ad un modello migliore che colga la relazione tra i dati.
In merito ai restanti grafici, si hanno dei chiari esempi in cui la retta di regressione non è uno strumento adeguato a modellare determinati tipi di relazione. Considerando, ad esempio, il grafico in basso a sinistra, si osserva che i punti si distribuiscono in maniera "verticale" concentrandosi su un intervallo limitato dei valori della variabile esplicativa.

[11] Osservazioni che influenzano notevolmente i risultati

La retta non coglie per nulla l'andamento dei dati.
I risultati ottenuti non hanno evidenziato alcuna relazione rilevante. Infatti i valori del coefficiente R^2 delle relazioni stimate sono risultati piuttosto bassi:

	R^2
NH4-As	0.30
Mn-Fe	0.26
As-Fe	0.18
Mn-As	0.01

3.3.3 Un esempio con Excel

In questa sezione, si fornisce un esempio dell'utilizzo di Excel (o la sua versione Open Office) per le analisi condotte in precedenza. Nello specifico si procede con l'analisi di correlazione tra Ferro e Ammonio, sfruttando le funzioni matematiche e statistiche già insite all'interno del software.
Innanzitutto viene calcolato il coefficiente di correlazione tra le variabili *NH4* e *Fe*. Ricordiamo la definizione

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{j=1}^{n}(y_j - \bar{y})^2}}.$$

Con Excel si può procedere in due modi: ricavarsi tutti gli ingredienti necessari per il calcolo del coefficiente, oppure utilizzare la formula *CORRELAZIONE()* già presente all'interno del programma.
Per la prima procedura è necessario calcolare la covarianza tra le variabili *NH4* e *Fe* e le rispettive deviazioni standard.
Per il calcolo della covarianza si usa il comando *COVARIANZA()*, mentre per calcolare le deviazioni standard si usa la funzione *DEV.ST()*.

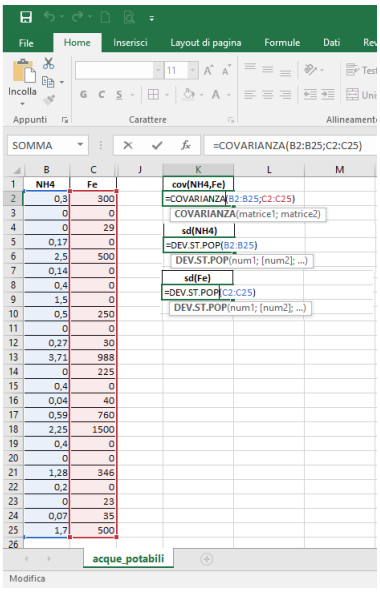


Figura 3.12: Calcolo della covarianza e delle deviazioni standard con Excel.

Infine per calcolare la correlazione è sufficiente applicare la formula (3.7).

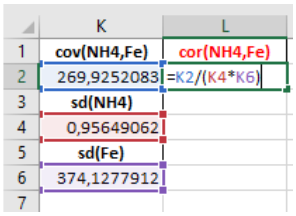


Figura 3.13: Calcolo della correlazione con Excel.

La seconda modalità di calcolo della correlazione comporta l'utilizzo della funzione *CORRELAZIONE()*.
Ovviamente, in entrambi i casi il valore della correlazione tra Ferro e Ammonio è risultato 0.75.
Excel, inoltre, fornisce funzioni in grado di costruire diversi tipi di grafico, tra cui quello di dispersione.

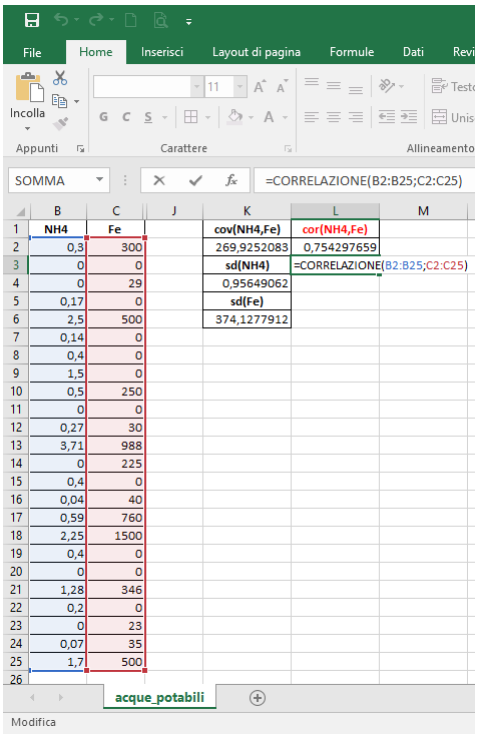


Figura 3.14: Calcolo della correlazione tramite la funzione di Excel.

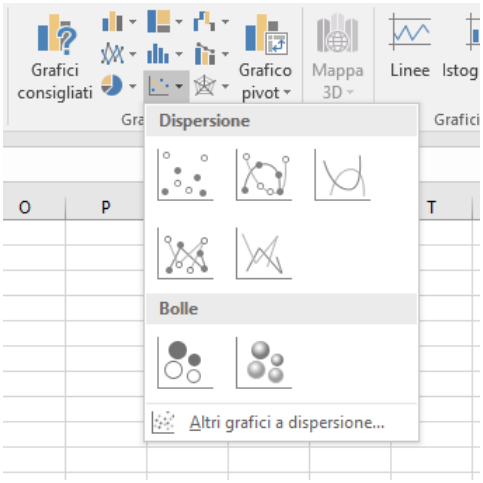


Figura 3.15: Funzione per la costruzione del grafico di dispersione.

Selezionando le serie dei dati inerenti alle variabili *Fe* e *NH4* si ottiene il grafico in Figura 3.17.

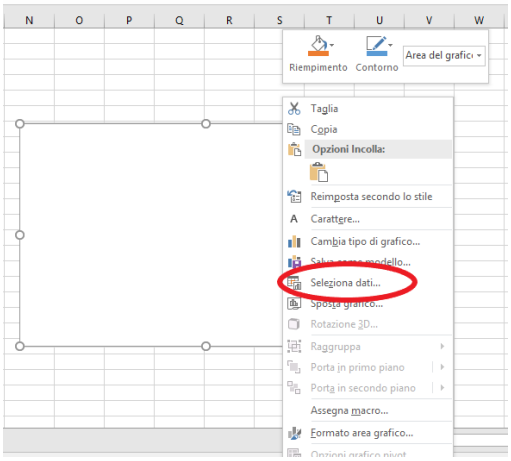


Figura 3.16: Funzione per la costruzione del grafico di dispersione.

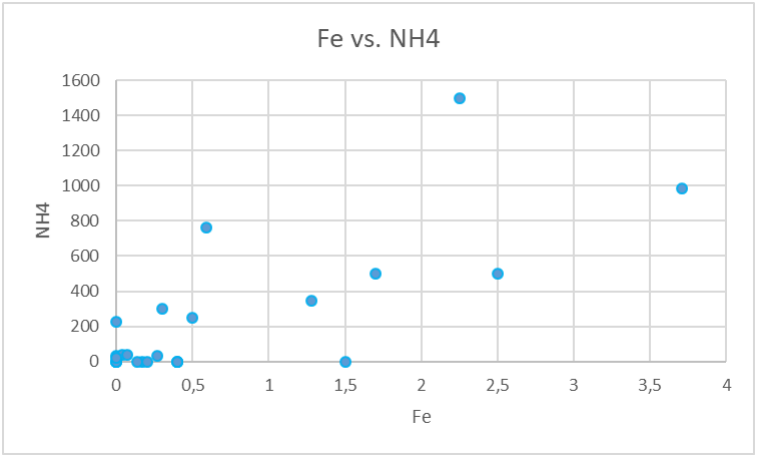
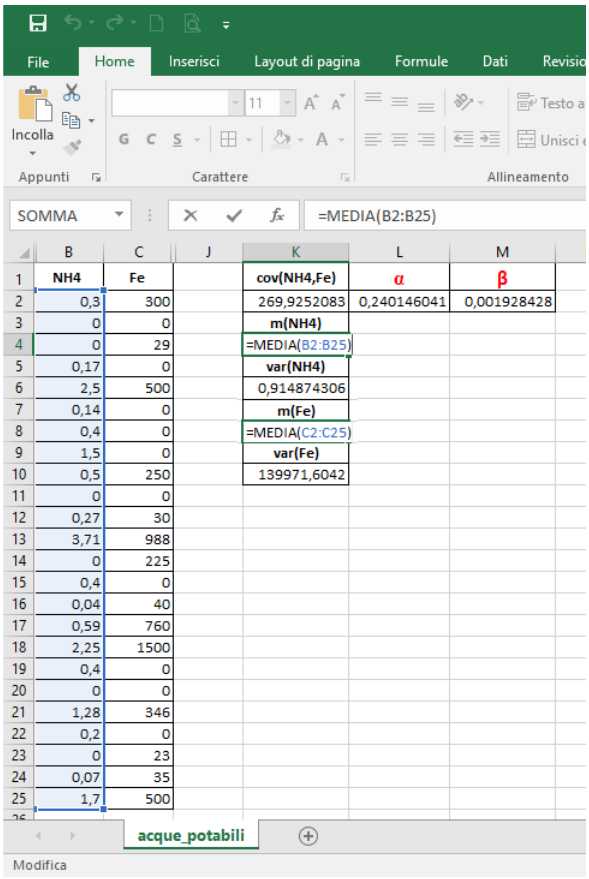


Figura 3.17: Grafico di dispersione Fe vs. NH4.

Per stimare il modello di regressione

$$NH4 = \alpha + \beta \times Fe + \epsilon$$

si applicano le formule (3.10) e (3.11) per la stima dei parametri α e β. Per calcolare le medie delle variabili *Fe* e *NH4* è sufficiente sfruttare la funzione *MEDIA()* presente all'interno di Excel.



	B	C	J	K	L	M
1	NH4	Fe		cov(NH4,Fe)	α	β
2	0,3	300		269,9252083	0,240146041	0,001928428
3	0	0		m(NH4)		
4	0	29		=MEDIA(B2:B25)		
5	0,17	0		var(NH4)		
6	2,5	500		0,914874306		
7	0,14	0		m(Fe)		
8	0,4	0		=MEDIA(C2:C25)		
9	1,5	0		var(Fe)		
10	0,5	250		139971,6042		
11	0	0				
12	0,27	30				
13	3,71	988				
14	0	225				
15	0,4	0				
16	0,04	40				
17	0,59	760				
18	2,25	1500				
19	0,4	0				
20	0	0				
21	1,28	346				
22	0,2	0				
23	0	23				
24	0,07	35				
25	1,7	500				

Figura 3.18: Calcolo delle medie con Excel.

Infine si applicano le formule per il calcolo dei parametri.

	K	L	M
1	cov(NH4,Fe)	α	β
2	269,9252083	0,240146041	=K2/K10
3	m(NH4)		
4	0,684166667		
5	var(NH4)		
6	0,914874306		
7	m(Fe)		
8	230,25		
9	var(Fe)		
10	139971,6042		

Figura 3.19: Calcolo di $\hat{\beta}$ con Excel.

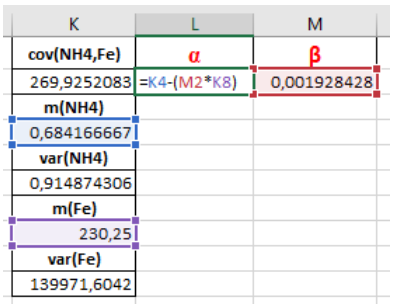


Figura 3.20: Calcolo di $\hat{\alpha}$ con Excel.

La retta stimata risulta:

$$N\hat{H}4 = 0.24 + 0.0019 \times Fe.$$

Excel mette a disposizione una funzione in grado di calcolare il coefficiente angolare della retta di regressione. Tale funzione si chiama *REGR.LIN()*.

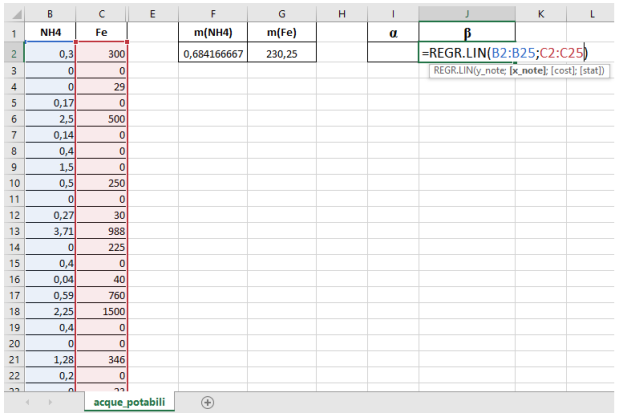


Figura 3.21: Calcolo di $\hat{\beta}$.

Per ottenere $\hat{\alpha}$ è sufficiente applicare la (3.11).

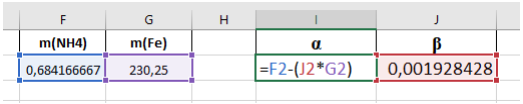


Figura 3.22: Calcolo di $\hat{\alpha}$.

Le stime ottenute coincidono con quelle calcolate in precedenza. Questa procedura risulta molto più veloce rispetto a quella descritta precedentemente, perciò è consigliabile utilizzarla per ottenere risultati in tempi più veloci.
Per disegnare la retta stimata sul grafico di dispersione, Excel mette a disposizione una funzione specifica.

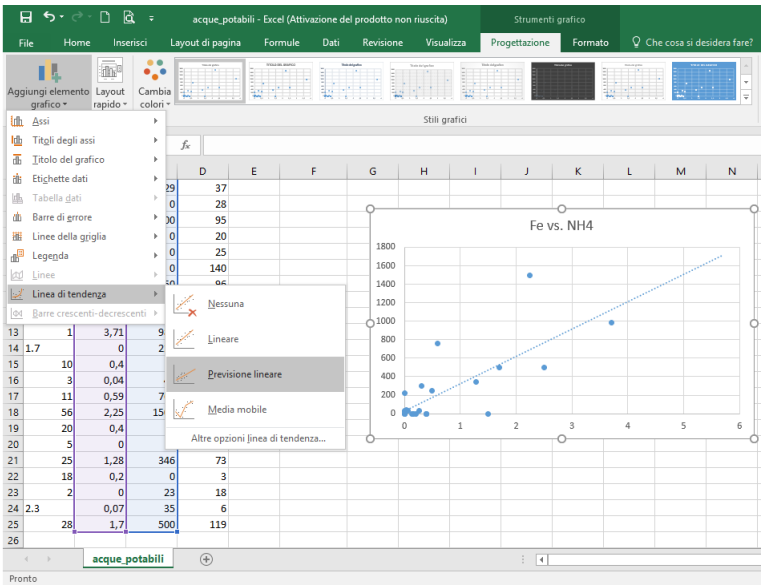


Figura 3.23: Funzione per il grafico della retta di regressione stimata.

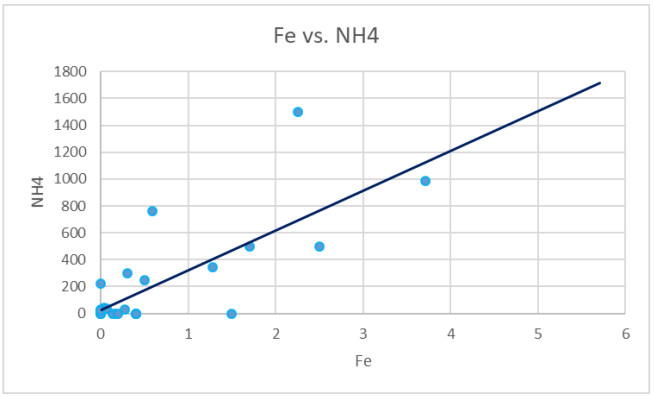


Figura 3.24: Retta di regressione stimata.

Questo dimostra come si possa utilizzare uno strumento informatico per produrre analisi statistiche in relazione ad un obiettivo funzionale. Una alternativa è l'uso del software R (https://www.r-project.org/). Si veda *Appendice A* per il codice R.
Concludendo si può affermare che l'applicazione appena descritta ha dimostrato l'interdisciplinarità della Statistica, utilizzando strumenti matematico-statistici per l'analisi di fenomeni reali relativi ad altri ambiti disciplinari, come la Chimica.
Nel prossimo capitolo si procede con la stesura di un altro modulo didattico che ha lo scopo di avvicinare gli studenti alla Statistica descrivendo un argomento intrigante ed innovativo: la *Sentiment Analysis*.

Capitolo 4

Modulo didattico: la *Sentiment Analysis* applicata a Twitter

In questo capitolo viene presentato un secondo modulo didattico, che affronta il tema della *Sentiment Analysis* applicata a Twitter. Lo scopo è mostrare come la Statistica possa essere utilizzata in svariati contesti, tra i quali l'analisi delle opinioni delle persone su un argomento predeterminato attraverso i *social networks*. Strumenti come i *social networks* (Facebook, Twitter,...) forniscono agli utenti della rete un punto d'incontro virtuale per scambiarsi messaggi, chattare, condividere foto e video, e tante altri servizi. Si sono diffusi in maniera rilevante nella società e, soprattutto, tra i ragazzi in età scolastica. Pertanto si è pensato che presentare un esempio in cui vengono utilizzati strumenti statistici per condurre analisi inerenti ad un *social network* possa incuriosire lo studente e catturare la sua attenzione. Nello specifico, nell'esempio considerato vengono scaricati $n = 600$ tweets[1] relativi a tre diverse marche di cellulare (I-phone, Samsung e Huawei) e poi classificati manualmente in tre categorie: *positivi*, *neutri* e *negativi*. Lo scopo è analizzare la soddisfazione dei consumatori, in relazione alle tre marche considerate, senza la necessità di intervistarli personalmente, ma solo "ascoltando" le opinioni che esprimono attraverso Twitter.

[1]Messaggio di testo avente una lunghezza non superiore a 140 caratteri, inviato a un sito Internet tramite instant messenger, e-mail o cellulare con lo scopo di comunicare informazioni in tempo reale.

Prima di vedere nel concreto l'esempio è necessario introdurre i concetti più rilevanti della *Sentiment Analysis*.

4.1 La *Sentiment Analysis*

La *Sentiment Analysis* (Ceron *et al.*, 2013, Liu, 2012) o *Analisi del Sentimento* (detta anche *Opinion Mining*) è una metodologia che si riferisce all'uso dell'elaborazione del linguaggio naturale, analisi testuale e linguistica-computazionale per identificare ed estrarre informazioni soggettive da diverse fonti. L'Analisi del Sentimento è ampiamente applicata per analizzare *social media* per una varietà di applicazioni, dal marketing al servizio clienti.
Gli aspetti positivi legati a questo tipo di analisi sono molteplici.

- Le opinioni degli utenti non vengono sollecitate dall'analista ma vengono catturate quando sono già state espresse, in maniera retrospettiva. Perciò non è necessario porre delle domande alle persone ma ci si limita ad "ascoltare" ciò che hanno da dire attraverso i social.
- Si possono condurre analisi in tempo reale, attraverso un monitoraggio continuo dei social ottenendo risultati in tempi brevi.
- I dati che vengono raccolti sono geo-localizzati, in modo da capire da dove provengono le informazioni raccolte.
- Si possono produrre analisi di tipo censuario, dato che molto spesso si ha a disposizione l'intera "popolazione" e non solo un campione.

Tuttavia la *Sentiment Analysis* presenta anche alcuni aspetti problematici.

- La popolazione sui *social networks* non è rappresentativa della popolazione demografica, in quanto solo persone di determinate fasce d'età possiedono uno o più *social media*.
- Come detto in precedenza, con questo tipo di analisi ci si limita solo ad "ascoltare" le opinioni perciò se gli utenti non trattano un determinato argomento non si possono avere informazioni in merito.

- Il linguaggio è in continua evoluzione e cambia a seconda dell'argomento e del contesto. Pertanto risulta complicato costruire un dizionario ontologico che cataloghi tutte le regole semantiche.

- La *Sentiment Analysis* lavora con *big data*, che possono presentarsi in quantità davvero elevata. Per *big data* si intendono raccolte di dataset sia molto grandi, ma anche così complessi da dover essere costretti ad utilizzare strumenti specifici per poterli trattare nelle fasi di acquisizione, gestione, analisi e per la visualizzazione dei dati di cui sono composti

4.1.1 La riduzione del testo in dato quantitativo: *lo Stemming*

Lo studio e l'analisi di opinioni, sentimenti, giudizi ed emozioni sono l'obiettivo della *Sentiment Analysis*, un'indagine di tipo statistico su documenti contenenti del testo, come i commenti Facebook o Twitter resa possibile dalle nuove potenzialità dei calcolatori e dalla nuova immensa disponibilità di dati che gli utenti del web generano spontaneamente ogni giorno.
Per poter analizzare le opinioni è necessario ridurre il testo in un dato quantitativo in modo da poter utilizzare un modello statistico per trattarlo.
Tutti i testi sono contenuti in un insieme, detto *corpus*, e vengono ridotti ad un gruppo di parole dette *stilemi* o *stem*. Gli stilemi possono essere formati da una singola parola (*unigram*) oppure se si vuole tenere conto di un determinato ordine possono essere costituiti da una coppia di parole (*bigram*) o una terna (*trigram*).
Il processo finalizzato alla riduzione dei testi in stilemi viene definto *stemming*. In seguito a questo procedimento si va a creare una matrice in cui nelle righe si trovano i testi, mentre nelle colonne gli stilemi. Nelle celle della matrice viene inserito il valore 1 se nel testo relativo compare quel stilema, 0 altrimenti.
Lo scopo è quello di trovare la distribuzione aggregata delle opinioni delle persone in merito a determinati argomenti. Nello specifico, in questo caso si

vogliono classificare le opinioni degli utenti in base ai loro tweets. Attraverso il software R (https://www.r-project.org/) si vanno a produrre delle analisi, prettamente descrittive, in cui le opinioni vengono discriminate in *positive*, *neutre* e *negative* (si veda *Appendice B* per il codice R).

4.1.2 L'analisi delle opinioni degli utenti di Twitter

Inizialmente è necessario scaricare i tweets dal sito ufficiale di Twitter (https://twitter.com) attraverso il software R. Dopo una prima fase di pulizia dei dati, in cui vengono rimosse dal tweet tutte le particelle grammaticali (congiunzioni, punteggiatura, articoli, preposizioni, suffissi, prefissi, ecc.), si procede con l'analisi delle opinioni. Sulla base di dizionari ontologici[2] esse vengono classificate nelle tre modalità definite in precedenza attraverso il *Sentiment Score*.

Cos'è il *Sentiment Score*?
Il *Sentiment Score* è una variabile qualitativa che assume le modalità $-1, 0, 1$ in base al sentimento con cui un'opinione è stata espressa nel documento, in questo caso nel tweet. Esiste un algoritmo che calcola tale indice servendosi dell'elenco di parole "positive" e "negative" prima citate.

Come si calcola il *Sentiment Score*?
L'algoritmo analizza uno per volta gli elementi del tweet e confronta gli stem contenuti nel testo con quelli presenti nelle due liste di parole. Gli stem vengono esaminati singolarmente e ad ognuno si assegna uno score positivo (+1) o negativo (-1) se viene individuato nel rispettivo dizionario ontologico. Nel caso di assenza di riscontro, lo score viene impostato a zero. Infine i punteggi (per tweet) vengono sommati:

[2]Due liste di parole italiane che esprimono una polarità, che può essere positiva o negativa.

- se $somma > 0$, significa che il tweet ha un "sentimento positivo" ($Score = +1$);
- se $somma = 0$, significa che il tweet ha un "sentimento neutro" ($Score = 0$);
- se $somma < 0$, significa che il tweet ha un "sentimento negativo" ($Score = -1$).

Nell'esempio delle marche dei cellulari si hanno $n = 600$ tweet, 200 per ogni marca, contenuti in tre *dataset* (formato Excel). I dati si ottengono "interrogando" Twitter in modo che restituisca tutti i tweet che contengono al loro interno gli *hashtag*[3] $\#iphone, \#samsung$ e $\#huawei$. Ogni *dataset* è composto da 7 variabili: *ID_tweet*, l'dentificativo del tweet; *stem_pos*, variabile quantitativa discreta che conta il numero di stem positivi contenuti nel tweet; *stem_neg*, variabile quantitativa discreta che conta il numero di stem negativi; *pt_pos*, variabile quantitativa discreta che descrive il punteggio degli stem positivi nel tweet; *pt_neg*, variabile quantitativa discreta che descrive il punteggio degli stem negativi nel tweet; *somma*, variabile quantitativa discreta che assume il valore della somma dei punteggi; *score*. Il *dataset* relativo al cellulare I-phone è:

	ID_tweet	stem_pos	stem_neg	pt_pos	pt_neg	somma	score
1	ip001	12	15	12	-15	-3	-1
2	ip002	7	7	7	-7	0	0
3	ip003	7	8	7	-8	-1	-1
	...						
	...						
200	ip200	14	1	14	-1	13	1

[3]Un *hashtag* è un tipo di etichetta (tag) utilizzato su alcuni servizi web e social networks come aggregatore tematico, la cui funzione è di rendere più facile per gli utenti trovare messaggi su un tema o contenuto specifico. Il termine *hashtag* può anche fare riferimento al simbolo cancelletto stesso, quando usato nel contesto di un hashtag.

Dal *dataset* si può osservare che in corrispondenza di valori negativi della variabile *somma* lo *score* assume valore -1; per valori positivi allo *score* viene associato il valore 1; per valori nulli assume valore 0. Si sottolinea, inoltre, che non compare la variabile che conta gli stem neutri, perchè a questi viene associato un punteggio nullo che non incide sulla variabile *somma*. I *dataset* delle altre due marche di cellulari sono analoghi.
Il prossimo passo consiste nel rappresentare graficamente le opinioni degli utenti. Trattandosi di variabili discrete i grafici più adatti a questo tipo di variabili sono il *diagramma a barre* e il *grafico a torta*.

4.1.3 Rappresentazione grafica delle opinioni

Il *diagramma a barre* è un grafico che viene utilizzato per la rappresentazione grafica di variabili discrete. Il grafico a barre è costituito da rettangoli di base uguale e altezza proporzionale alla frequenza assoluta o relativa con cui si presentano le modalità, chiamati *barre*. A differenza dell'istogramma, i rettangoli sono separati gli uni dagli altri per sottolineare l'assenza di continuità nella variabile. Il numero di barre rappresentate è pari al numero di modalità della variabile considerata.

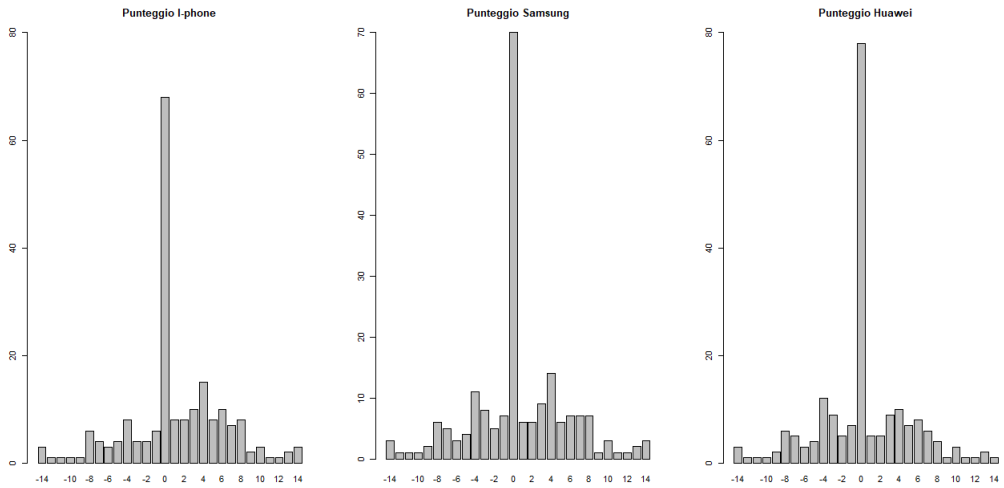


Figura 4.1: Diagramma a barre della somma dei punteggi per le tre marche.

Trattandosi di variabili numeriche è possibile costruire i box-plot relativi alle tre variabili:

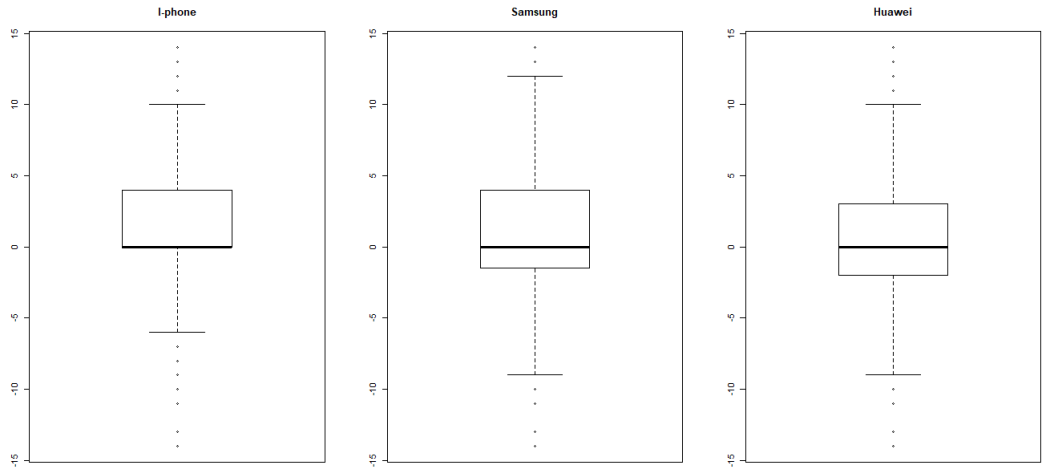


Figura 4.2: Box-plot della somma dei punteggi per le tre marche.

Dalle Figure 4.1 e 4.2 si osserva che tutte le distribuzioni dei punteggi sono centrate sullo 0, osservando un'alta frequenza per tale modalità (ossia la moda). Quindi si può affermare che le distribuzioni risultano sostanzialmente simmetriche.
La *moda* è la modalità (o la classe di modalità) caratterizzata dalla massima frequenza, ossia il valore che compare più frequentemente.
In questo caso la moda coincide con la modalità 0. Inoltre si osservano alcuni valori anomali per tutte le distribuzioni, soprattutto per l'I-phone. Vengono riportate, di seguito, medie, mediane, sd e IQR delle tre variabili:

	Media	**Mediana**	**Dev. Std.**	**Scarto Inter.**
I-phone	1.04	0	5.14	4.00
Samsung	0.54	0	5.16	5.25
Huawei	0.17	0	4.88	5.00

Tutte queste considerazioni potrebbero trarre in inganno e portare alla conclusione che le opinioni dei consumatori sono sostanzialmente neutre. Tuttavia è necessario ricordare che le modalità di questa variabile vengono raggruppate per costruire lo *score*, l'indice che permette di trarre conclusioni

sulle opinioni. Certamente si può osservare che Samsung e Huawei hanno, in proporzione, un numero più elevato di tweet con punteggio neutro rispetto all'I-phone. Dunque per comprendere, effettivamente, quale sia la soddisfazione degli utenti di Twitter rispetto alle tre marche è necessario anlizzare lo *score*. In questo caso sono stati considerati $n = 200$ tweets per ogni marca (in totale 600) e sono state rilevate le seguenti frequenze relative:

	f_{pos}	f_{neutro}	f_{neg}
I-phone	0.23	0.34	0.43
Samsung	0.285	0.35	0.365
Huawei	0.295	0.39	0.315

Attraverso le frequenze relative è possibile costruire i diagrammi a barre delle opinioni per ogni marca di cellulare.

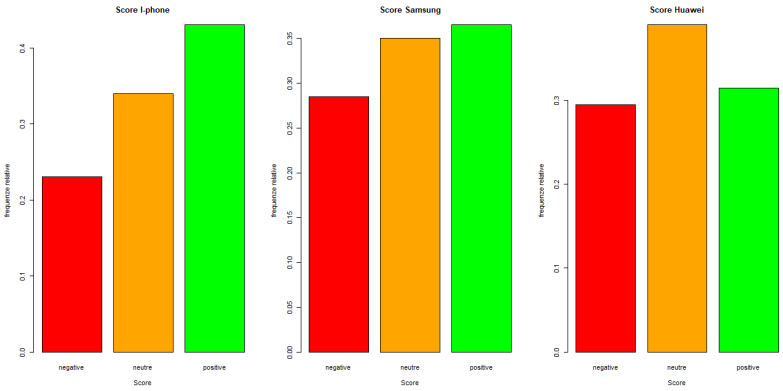


Figura 4.3: Diagramma a barre dello Score.

Dalla Figura 4.3 si può osservare che le distribuzioni delle opinioni di I-phone e Samsung sono spostate verso destra, riscontrando un'elevata frequenza della modalità *opinioni-positive*, segnale che i consumatori sono soddisfatti dei prodotti inerenti a tali marche. Tuttavia per la marca Huawei il diagramma a barre mostra una frequenza molto elevata della modalità *opinioni-neutre* ad indicare una neutralità delle opinioni degli utenti. Quindi le persone non sono né soddisfatte né insoddisfatte dei prodotti dell'azienda. Probabilmente sarebbe ideale promuovere una politica mirata ad aumentare il grado di

soddisfazione dei clienti, migliorando la qualità dei prodotti e dei servizi.A conferma di quanto detto, la moda della variabile Score è la modalità +1 (*opinioni-positive*) per I-phone e Samsung, con frequenza, rispettivamente, 43% e 36.5%. Coerentemente con quanto dedotto dal grafico, gli utenti esprimono prevalentemente opinioni positive sul social in relazione a queste marche. Mentre per Huawei si riscontra che la modalità con maggiore frequenza è *opinioni-neutre*, con frequenza pari al 39%.
Oltre ai grafici a barre esiste un'altra rappresentazione grafica delle variabili qualitative, ossia i *diagrammi a torta*. Un diagramma a torta viene costruito dividendo un cerchio in spicchi le cui ampiezze angolari sono proporzionali alle frequenze. Come per l'istogramma, le aree sono proporzionali alle frequenze. In questo caso si vanno a rappresentare per ogni modalità (positive, neutre e negative), le percentuali di ciascuna marca relative a quella modalità attraverso i diagrammi a torta.

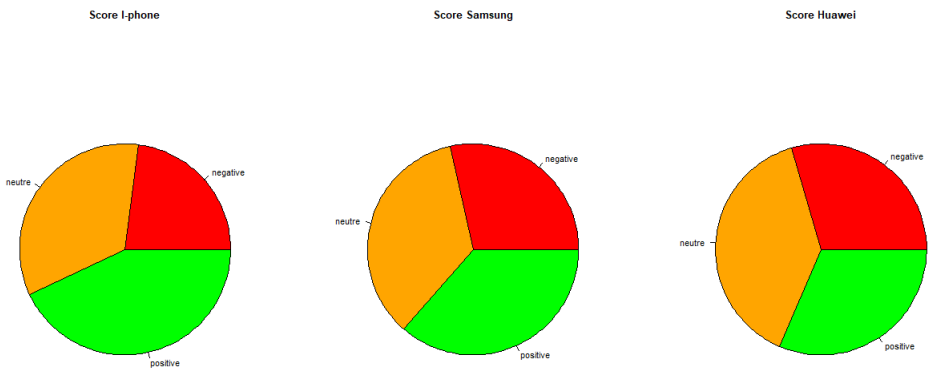


Figura 4.4: Diagramma a torta dello Score.

Dalla Figura 4.4 si può osservare che in proporzione la marca di cellulare che possiede la percentuale più elevata di opinioni negative è il Huawei, infatti il suo "spicchio" nel grafico ha un'area maggiore rispetto alle altre due. Come detto in precedenza, l'azienda Huawei dovrebbe pensare ad una poli-

tica che catturi più consensi da parte dei consumatori incrementando, così, la soddisfazione rispetto ai suoi prodotti. Ragionando analogamente si può dire che tra le tre marche, il tipo di cellulare maggiormente apprezzato dagli utenti è l'I-phone, data la sua alta percentuale di opinioni positive rispetto alle altre due. L'azienda dell'I-phone (Apple) punta molto sull'innovazione e ha saputo dare, nell'ultimo decennio, una svolta nel mondo della tecnologia diffondendo una vera e propria moda nella società, che porta i consumatori ad appassionarsi ai suoi prodotti.

4.1.4 Le analisi con Excel

In questa sezione si descrive l'uso di Excel per condurre le stesse analisi viste in precedenza. Il *dataset* è contenuto in un file Excel pronto per l'utilizzo.

	A	B	C	D	E	F	G
1	ID_tweet	stem_pos	stem_neg	pt_pos	pt_neg	somma	score
2	ip001	12	15	12	-15	-3	-1
3	ip002	1	10	1	-10	-9	-1
4	ip003	7	8	7	-8	-1	-1
5	ip004	10	15	10	-15	-5	-1
6	ip005	4	10	4	-10	-6	-1
7	ip006	2	7	2	-7	-5	-1
8	ip007	11	13	11	-13	-2	-1
9	ip008	8	12	8	-12	-4	-1

Figura 4.5: Spezzone del dataset in Excel relativo all'I-phone.

Inizialmente si procede con il calcolo delle *frequenze assolute* della variabile *sent* che rappresenta lo *Score*, attraverso la funzione di Excel *CONTA.SE(intervallo, criterio)*. Grazie a questa funzione si possono contare tutte le righe del dataset che soddisfano il criterio specificato.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID_tweet	stem_pos	stem_neg	pt_pos	pt_neg	somma	score				Score		
2	ip001	12	15	12	-15	-3	-1				-1	0	1
3	ip002	1	10	1	-10	-9	-1			frequenze assolute	=CONTA.SE(G2:G201;-1)		
4	ip003	7	8	7	-8	-1	-1			frequenze relative	CONTA.SE(intervallo; criterio)		
5	ip004	10	15	10	-15	-5	-1						
6	ip005	4	10	4	-10	-6	-1						
7	ip006	2	7	2	-7	-5	-1						

Figura 4.6: Calcolo frequenze assolute.

Dalla Figura 4.6 si vede l'applicazione della funzione *CONTA.SE()* per il calcolo delle *frequenze assolute* della modalità -1 dello *Score*. Viene selezionata la colonna della variabile *score* e viene specificato il criterio -1. Così facendo Excel conta tutte le celle della colonna che assumono il valore -1. Analogamente si procede per le altre due modalità.
Successivamente si vanno a calcolare le *frequenze relative*, dividendo quelle assolute per la numerosità dei dati ($n = 200$).

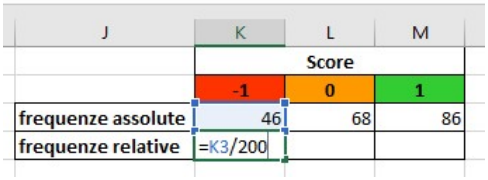


Figura 4.7: Calcolo frequenze relative.

Una volta calcolate le frequenze relative è possibile costruire il diagramma a barre. Dopo aver selezionato i valori da rappresentare è sufficiente andare su *Inserisci* → *Grafici* e cliccare su *Inserisci grafico statistico* → *Altri grafici statistici*.

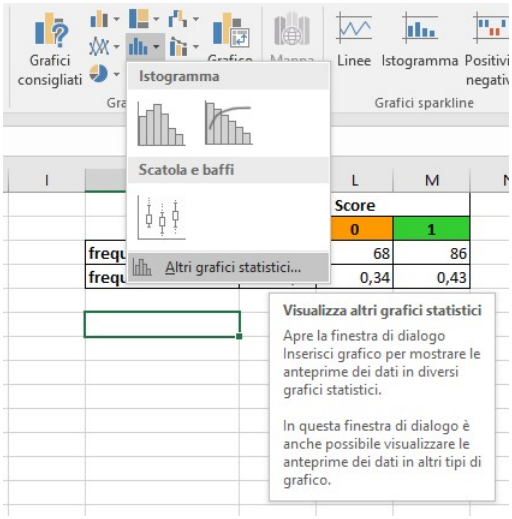


Figura 4.8: Inserimento di un grafico statistico.

Appare, così, una finestra che permette di scegliere il tipo di grafico da rappresentare, in questo caso il diagramma a barre.

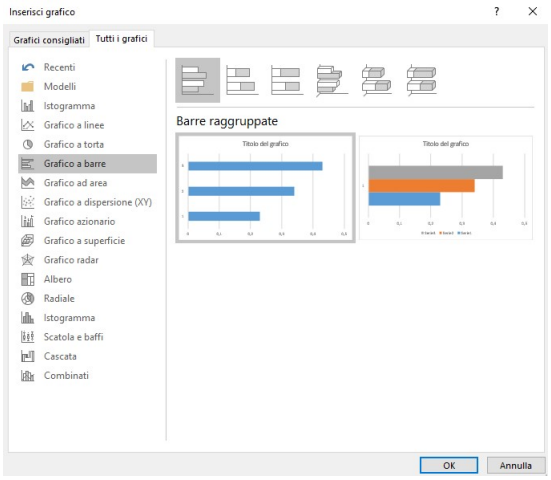


Figura 4.9: Inserimento di un diagramma a barre.

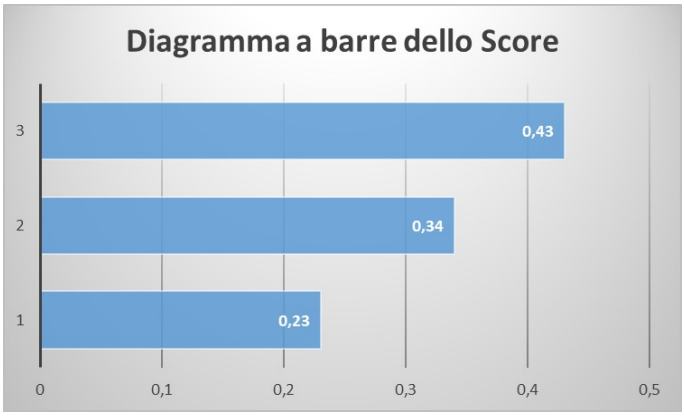


Figura 4.10: Diagramma a barre dello Score.

Le barre, in questo caso, sono posizionate orizzontalmente e le modalità sono codificate con i numeri $1 = \textit{negative}, 2 = \textit{neutre}, 3 = \textit{positive}$. Tuttavia non cambia nulla in termini di conclusioni: la modalità $3 = \textit{positive}$ è quella che si presenta con frequenza maggiore, indicando una sostanziale positività delle opinioni degli utenti di Twitter.
Per costruire il diagramma a torta i passi da seguire sono i medesimi eccetto

per la scelta del grafico. Questa volta è sufficiente selezionare nella finestra in Figura 4.8 la voce *Grafico a torta*.

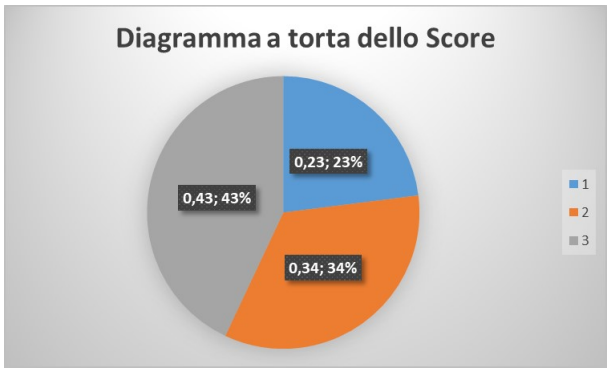


Figura 4.11: Diagramma a torta dello Score.

Analogamente si procede per la costruzione dei grafici con i dati degli altri due *dataset*.

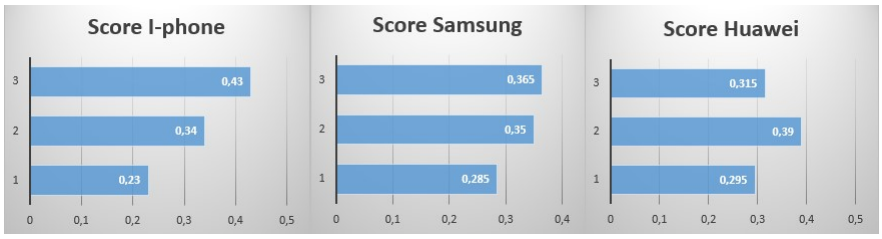


Figura 4.12: Diagramma a barre dello Score di I-phone, Samsung, Huawei.

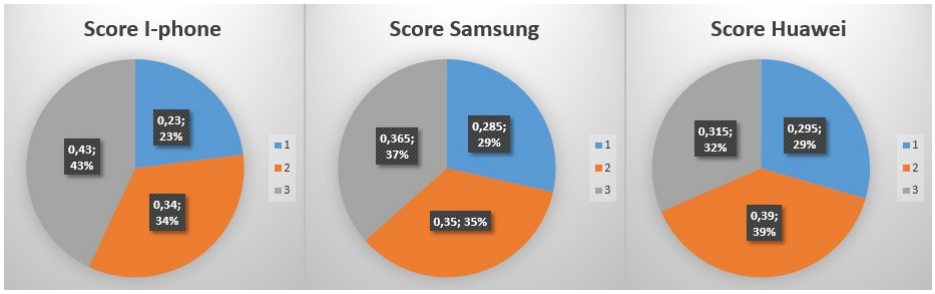


Figura 4.13: Diagramma a torta dello Score di I-phone, Samsung, Huawei.

Questo modulo serve a dimostrare come nozioni elementari di Statistica, che vengono trattate in classe, possano servire per affrontare tematiche relative a problemi concreti e interessanti per i ragazzi. La speranza è che esempi di questo tipo possano intrigare ed appassionare gli studenti in modo che si avvicinino alle Scienze Statistiche e intraprendano un percorso inerente una volta terminati gli studi nelle scuole superiori. Inoltre è anche un modo per far comprendere ai docenti di Matematica l'importanza di questa disciplina, sottolineando la vastità di applicazioni in cui la Statistica risulta uno strumento utile ed efficiente per risolvere problemi reali.

Capitolo 5

"Facciamo Statistica a scuola"

5.1 L'incontro con gli studenti dell'*I.T.I.S. E. Fermi* di Bassano del Grappa

In questo capitolo si commenta l'esperienza svoltasi nel *Dicembre 2017* con gli studenti delle classi 4° e 5° dell'indirizzo chimico-biologico dell'ITIS *E. Fermi* di Bassano del Grappa.
L'incontro si è articolato in due parti distinte. Nella prima parte sono stati presentati ai ragazzi i due esempi trattati in precedenza nei moduli didattici:

- *Monitoriaggio delle acque potabili del Veneto* per illustrare le tematiche relative a correlazione e regressione;
- *La Sentiment Analysis applicata a Twitter* per intrigare gli studenti con un argomento innovativo e di attualità.

Nella seconda parte, invece, sono stati presentati i corsi di laurea in Statistica in modo da rendere coscienti gli studenti che un percorso di studio del genere risulta necessario per affrontare problemi e tematiche affini a quelle illustrate nella prima parte dell'incontro.
Al termine delle due presentazioni è stato sottoposto un questionario di gradimento agli studenti, in modo da avere un riscontro da parte loro sull'esperienza svolta.

Sono stati compilati 39 questionari che, successivamente, sono stati analizzati per avere una panoramica complessiva sull'esperienza con i ragazzi.
Sono state considerate ed analizzate a livello descrittivo 10 variabili qualitative:

- GIUDIZIO: variabile qualitativa che descrive il giudizio complessivo dello studente in merito all'incontro;
- CONOSCENZA_PRE: variabile qualitativa che descrive quanto le conoscenze preliminari degli studenti siano state utili per seguire l'incontro;
- APPROF_TEMATICHE: variabile qualitativa che descrive, a giudizio dello studente, quanto sono state approfondite le tematiche durante l'incontro;
- METODOLOGIA: variabile qualitativa che descrive il giudizio dello studente rispetto alla metodologia utilizzata durante l'incontro;
- ASPETTATIVE: variabile qualitativa che descrive le aspettative degli studenti rispetto all'incontro;
- UTILITA: variabile qualitativa che descrive l'utilità dell'incontro secondo gli studenti.

Tutte le variabili elencate sono caratterizzate da 5 modalità:

- 1 = MOLTO NEGATIVO
- 2 = NEGATIVO
- 3 = NEUTRO
- 4 = POSITIVO
- 5 = MOLTO POSITIVO

Inoltre sono state considerate altre due variabili qualitative:

- SESSO: indica il sesso dello studente (M o F);
- CLASSE: indica la classe frequentata dallo studente (4 o 5).

5.1.1 Analisi complessive del gradimento degli studenti

Vengono riportate di seguito le analisi di sintesi relative al gradimento complessivo degli studenti.

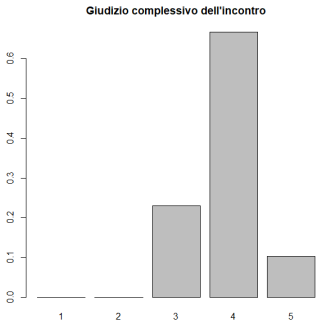


Figura 5.1: Giudizio complessivo dell'incontro.

Dal diagramma a barre in Figura 5.1 si osserva che gli studenti hanno espresso dei giudizi positivi nei confronti dell'incontro tenuto a scuola; infatti nessuno ha scelto le modalità 1 (Molto negativo) o 2 (Negativo). La moda di questa variabile corrisponde a 4 (Positivo) ad indicare complessivamente un riscontro positivo da parte degli studenti. La media, in questo caso, è pari 3.87. L'incontro ha suscitato interesse raggiungendo efficientemente l'obiettivo prefissato: avvicinare gli studenti alla Statistica.

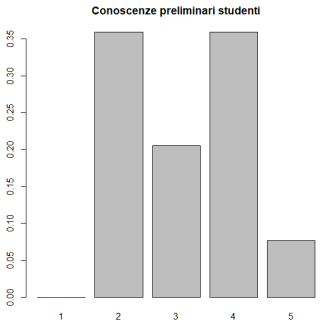


Figura 5.2: Conoscenze preliminari studenti.

Dalla Figura 5.2 si osserva che si vanno a creare sostanzialmente due gruppi. Da una parte ci sono gli studenti che hanno ritenuto di non avere avuto delle conoscenze adeguate per seguire l'incontro e dall'altra coloro che possedevano già le nozioni utili. Alla base di questa differenziazione sta la classe frequentata dallo studente: infatti gli studenti della classe 4° non avevano ancora affrontato le tematiche relative a correlazione e regressione, mentre i ragazzi di 5° erano consapevoli di ciò che si stava trattando durante l'incontro. La media risulta 3.15 ad indicare una sostanziale neutralità.

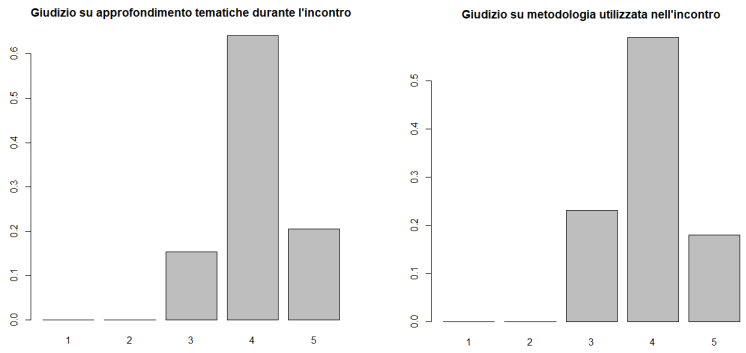


Figura 5.3: Approfondimento delle tematiche e Metodologia della spiegazione.

Dalla Figura 5.3 si può osservare che i giudizi risultano prevalentemente positivi e la moda corrisponde a 4 (Positivo). Le medie di entrambe le variabili confermano quanto detto, in quanto assumono i valori 4.05 per la prima e 3.95 per la seconda. Quindi dal punto di vista degli studenti le tematiche affrontate durante l'incontro sono state approfondite in maniera adeguata, considerando la loro preparazione di base e l'interesse rispetto alla materia. Inoltre hanno espresso un parere positivo in merito alla metodologia utilizzata nella spiegazione delle nozioni e all'organizzazione dell'incontro in termini di materiale e strumenti per l'acquisizione delle conoscenze.

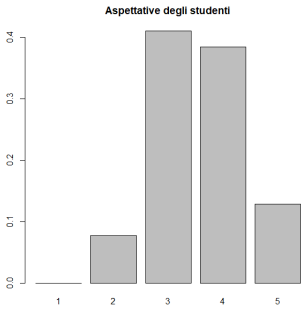


Figura 5.4: Aspettative degli studenti.

Dal grafico in Figura 5.4 si osserva che la moda corrisponde alla modalità 3 (Neutro), mentre la media è pari a 3.56 ad indicare una sostanziale neutralità rispetto alle aspettative che gli studenti avevano prima dell'incontro (circa il 40% degli studenti). Tuttavia si osserva un'alta frequenza per la modalità 4 (circa il 38% degli studenti), quindi si può dire che per molti studenti l'incontro è stato all'altezza delle aspettative.
In merito all'utilità dell'incontro (Figura 5.5), secondo il giudizio dei ragazzi, si può affermare che per molti di loro l'esperienza è risultata utile; infatti la moda corrisponde a 4 (Positivo), mentre la media è pari a 3.92. In generale i giudizi sono stati positivi, dato che la maggior parte delle frequenze registrate sono relative alle modalità 3 (Neutro), 4 (Positivo) e 5 (Molto Positivo).

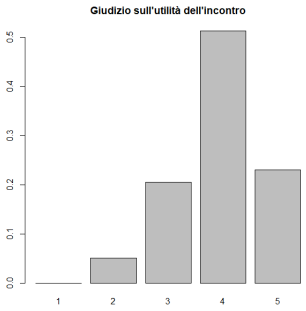


Figura 5.5: Giudizio sull'utilità dell'incontro.

	Moda	**Media**
Giudizio complessivo	4	3.87
Conoscenze preliminari	2	3.15
Approfondimento tematiche	4	4.05
Metodologia	4	3.95
Aspettative	3	3.56
Utilità	4	3.92

Traendo delle conclusioni generali, si può affermare che gli studenti hanno mosso dei giudizi prevalentemente positivi nei confronti dell'esperienza svolta a scuola. Hanno mostrato interesse e hanno apprezzato l'incontro nel suo complesso, ad indicare come questi tipi di progetti siano utili per divulgare la Statistica nelle scuole. La divulgazione è fondamentale per avvicinare gli studenti alla materia e rendere i docenti consapevoli di quanto sia importante dare maggiore peso alla Statistica nel percorso di studio dei ragazzi.
Di seguito vengono condotte le analisi differenziando i ragazzi secondo il sesso (SESSO) e la classe frequentata (CLASSE) per comprendere se vi siano differenze significative tra i gruppi esaminati.

5.1.2 Analisi del gradimento degli studenti rispetto al sesso

In questa sezione si eseguono le stesse analisi viste in precedenza, differenziando i ragazzi rispetto alla variabile SESSO. In totale gli studenti erano $n = 39$ contando $n_F = 14$ ragazze (circa il 36%) e $n_M = 25$ ragazzi (circa il 64%). Per confrontare le distribuzioni delle variabili nei due gruppi viene applicato il *test di Mann-Whitney* (o *test della somma dei ranghi*) per capire se vi sono differenze significative tra le distribuzioni nei due gruppi.

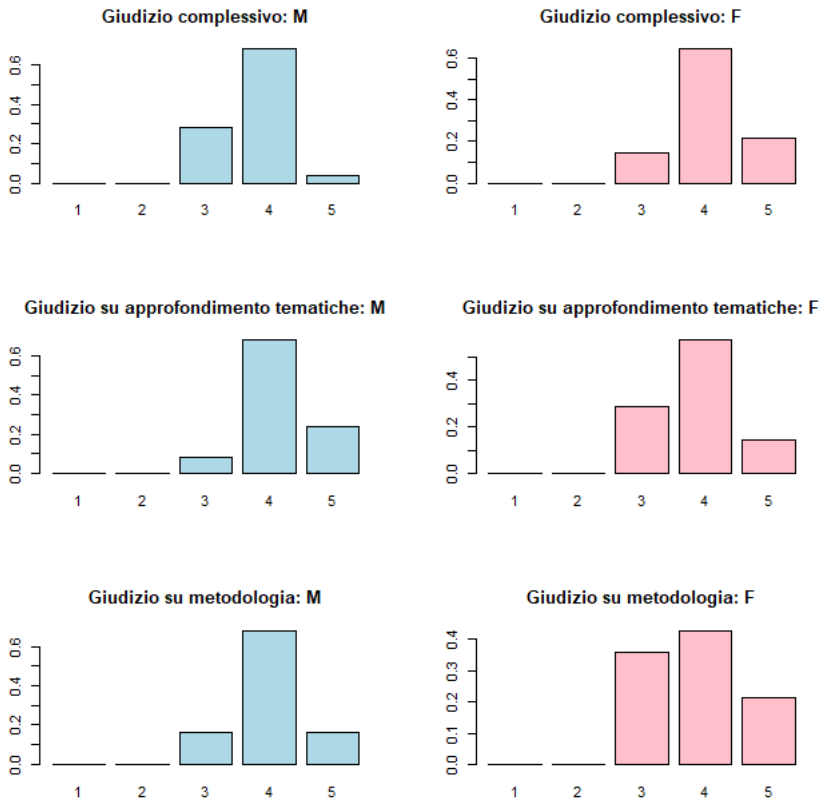


Figura 5.6: Giudizio sull'incontro, approfondimento tematiche e metodologia: M vs F.

	$\textbf{Moda}_M$	$\textbf{Moda}_F$	$\textbf{Media}_M$	$\textbf{Media}_F$
Giudizio complessivo	4	4	3.76	4.07
Approfondimento tematiche	4	4	4.16	3.86
Metodologia	4	4	4.00	3.86

Dalla Figura 5.6 si osservano distribuzioni simili in merito alle tre variabili considerate. Non si sottolineano grandi differenze tra maschi e femmine. Entrambi i gruppi hanno manifestato un giudizio positivo rispetto all'incontro e le modalità con sui è stato sviluppato. Infatti la moda per tutte le variabili corrisponde con la modalità 4 (POSITIVO) indipendentemente dal sesso degli studenti. Questo sta a significare che sono stati trattati temi che hanno colpito sia ragazzi che ragazze, indicando che la Statistica è una materia

mirata agli studenti di entrambi i sessi, a differenze di altri percorsi di studi che vedono la partecipazione di studenti prevalentemente di un solo sesso. Di seguito vengono riportati i valori osservati della statistica W e il relativo *p-value* ottenuti con il *test di Mann-Whitney* per il confronto delle distribuzioni delle variabili nei due gruppi (per ogni variabile considerata):

Variabile	**Statistica W**	**p-value**
Giudizio complessivo	130	0.11
Approfondimento tematiche	218	0.14
Metodologia	196	0.49

Tutti i test presentano un *p-value* > 0.05, ad indicare che non vi sono differenze significative tra i due gruppi, confermando quanto dedotto dai grafici in Figura 5.6.

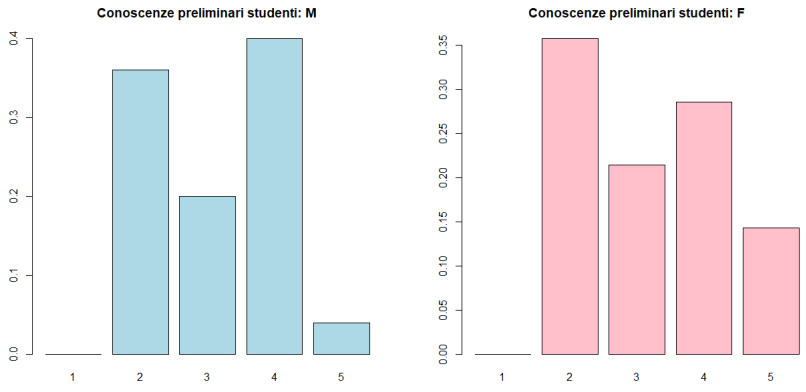


Figura 5.7: Conoscenze preliminari: M vs F.

Dalla Figura 5.7 si evince che per le ragazze, la moda coincide con la modalità 2 (NEGATIVO) mentre per i ragazzi con la modalità 4 (POSITIVO). Questo potrebbe indicare che rispetto agli studenti di sesso maschile, le studentesse non possedevano le conoscenze adeguate per affrontare in modo adeguato l'incontro. Inoltre, le medie delle due variabili sono pressochè uguali: $\bar{x_M} = 3.12$ e $\bar{x_F} = 3.21$. Tuttavia se si vanno a vedere i risultati

del test sui ranghi non si osservano differenze significative tra i due gruppi ($p\text{-}value > 0.05$).

	Variabile	**Statistica W**	**p-value**
	Approfondimento tematiche	168	0.83

Nel complesso, gli studenti di entrambi i sessi sono rimasti soddisfatti in maniera eguale dall'intervento, rispetto le proprie aspettative. Infatti il test sottolinea l'assenza di differenze tra i due gruppi.
Infine, nel complesso ragazzi e ragazze hanno manifestato lo stesso comportamento nell'esprimere la loro opinione sull'utilità dell'incontro. La maggior parte degli studenti ha ritenuto utile in modo positivo progetti di questo tipo; infatti la moda per entrambi i gruppi è pari a 4 (Figura 5.8).

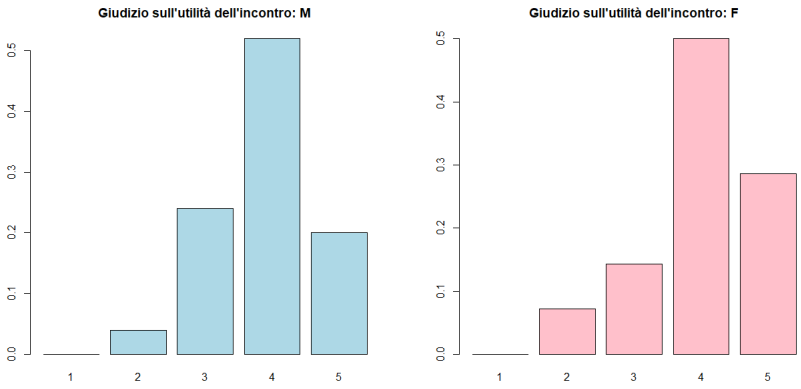


Figura 5.8: Giudizio sull'utilità dell'incontro: M vs F.

	Moda$_M$	**Moda$_F$**	**Media$_M$**	**Media$_F$**
Utilità	4	4	3.88	4.00

I risultati del test hanno confermato quanto detto: non si riscontrano differenze significative tra i due gruppi ($p\text{-}value > 0.05$).

Variabile	**Statistica W**	**p-value**
Utilità	157	0.57

Ora si riportano i risultati delle stesse analisi differenziando gli studenti in base alla classe frequentata, per vedere se vi sono differenze dovute alla diversa preparazione dei ragazzi.

5.1.3 Analisi del gradimento degli studenti rispetto alla classe frequentata

In questa sezione le analisi viste in precedenza sono state condotte differenziando i dati rispetto alla variabile CLASSE. In totale gli studenti erano $n = 39$, con $n_4 = 20$ studenti di classe 4° e $n_5 = 19$ di classe 5°.

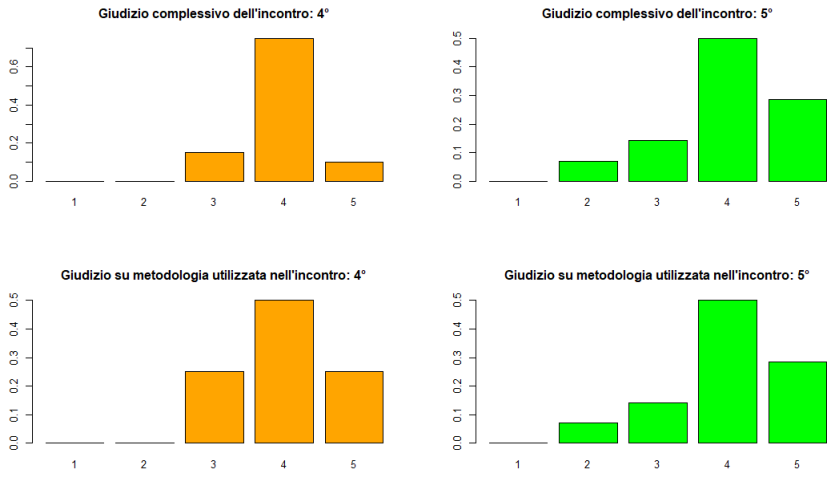


Figura 5.9: Giudizio complessivo e sulla metodologia: 4° vs 5°.

	$\mathbf{Moda}_{IV}$	$\mathbf{Moda}_{V}$	$\mathbf{Media}_{IV}$	$\mathbf{Media}_{V}$
Giudizio complessivo	4	4	3.95	3.79
Metodologia	4	4	4.00	3.89

Dalla Figura 5.9 si osservano distribuzioni simili delle due variabili considerate. Non si sottolineano grandi differenze tra studenti di classe 4° e 5°. Entrambi i gruppi hanno manifestato un giudizio positivo rispetto all'incontro e le modalità con sui è stato sviluppato. Infatti la moda per tutte le

variabili corrisponde con la modalità 4 (POSITIVO) e non vi sono giudizi negativi (al di sotto del punteggio 3) indipendentemente dalla classe frequentata. Questo sta a significare che sono state trattate temi che hanno saputo cogliere l'attenzione di ragazzi frequentanti classi diverse, nonostante al momento dell'incontro possedessero nozioni differenti (ad esempio i ragazzi di 4° non avevano ancora trattato gli argomenti inerenti a correlazione e regressione). A conferma di quanto detto si applica il test basato sui ranghi, i quali non evidenziano differenze significative tra i due gruppi ($p\text{-}value > 0.05$).

Variabile	**Statistica W**	**p-value**
Giudizio complessivo	217.5	0.35
Metodologia	205	0.63

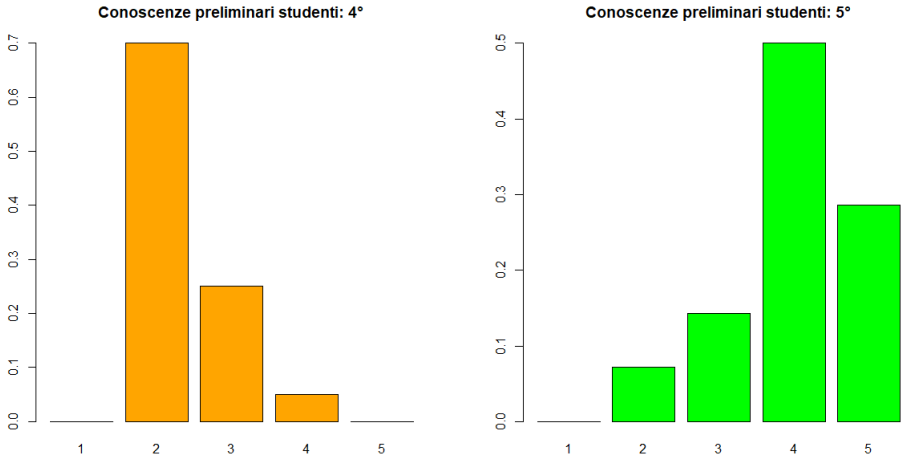


Figura 5.10: Conoscenze preliminari: 4° vs 5°.

	Moda$_{IV}$	**Moda**$_{V}$	**Media**$_{IV}$	**Media**$_{V}$
Conoscenze preliminari	2	4	2.35	4.00

Dalla Figura 5.10 si osserva una evidente differenza tra i due gruppi in termini di conoscenze preliminari. La moda per i dati relativi alla classe 4° è pari a 2 (NEGATIVO) mentre per la classe 5° è 4 (POSITIVO). Infatti i ragazzi di 4° al momento dell'incontro non possedevano le nozioni legate alle tematiche

di correlazione e regressione, che in generale vengono trattate nel secondo quadrimestre (periodo Marzo-Giugno). Questo ha reso complicato per loro seguire in modo efficiente l'intero incontro, tuttavia può essere stato recepito come uno stimolo per seguire con più interesse le lezioni in classe relative a queste tematiche. Per quanto riguarda i ragazzi di 5°, le conoscenze che possedevano sono risultate adeguate per partecipare in modo collaborativo all'incontro. Trattandosi di tematiche viste l'anno precedente, hanno potuto avere una contestualizzazione pratica di strumenti di cui avevano solo una concezione teorica.
A conferma di ciò si applica il *test di Mann-Whitney* e si osserva un valore del *p-value* < 0.05 ad indicare una differenza significativa tra i due gruppi considerati.

Variabile	**Statistica W**	**p-value**
Conoscenze preliminari	17	< 0.001

Come si osserva dalla Figura 5.11 si può affermare che i ragazzi di entrambe le classi hanno ritenuto che le tematiche affrontate durante l'incontro siano state approfondite in maniera adeguata. Infatti la moda per entrambi i gruppi risulta 4 (POSITIVO). Questo sta a significare che, nonostante le scarse conoscenze preliminari, anche gli studenti di 4° hanno apprezzato gli argomenti trattati durante l'intervento. Il test di *Mann-Whitney* non evidenzia differenze tra le distribuzioni:

Variabile	**Statistica W**	**p-value**
Approfondimento tematiche	191.5	0.96

	Moda$_{IV}$	**Moda**$_{V}$	**Media**$_{IV}$	**Media**$_{V}$
Approfondimento tematiche	4	4	4.05	4.05

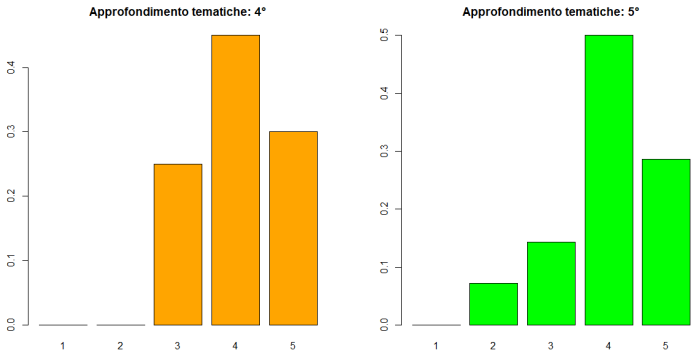


Figura 5.11: Approfondimento tematiche: 4° vs 5°.

Come ci si aspettava, i ragazzi di 5° sembrano essere stati maggiormente soddisfatti dell'incontro rispetto ai ragazzi di 4° che si sono mantenuti sostanzialmente neutrali (Figura 5.12). Tuttavia gli studenti di classe 4° sono stati rimasti colpiti in quanto hanno potuto vedere l'applicazione di strumenti fino ad allora sconosciuti. Si trattava di novità assolute estranee alle competenze da loro possedute, che hanno saputo catturare il loro interesse rispetto a quello degli altri studenti che già conoscevano questo tipo di strumenti.

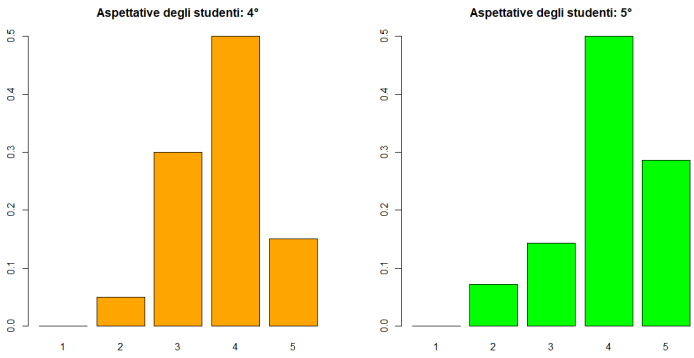


Figura 5.12: Aspettative: 4° vs 5°.

Tuttavia il test non riscontra una differenza significativa tra i gruppi:

Variabile	**Statistica W**	**p-value**
Aspettative	242	0.12

Dalla Figura 5.13, sembrerebbe che i ragazzi di classe 4° abbiano ritenuto l'incontro maggiormente utile rispetto ai ragazzi di classe 5°.

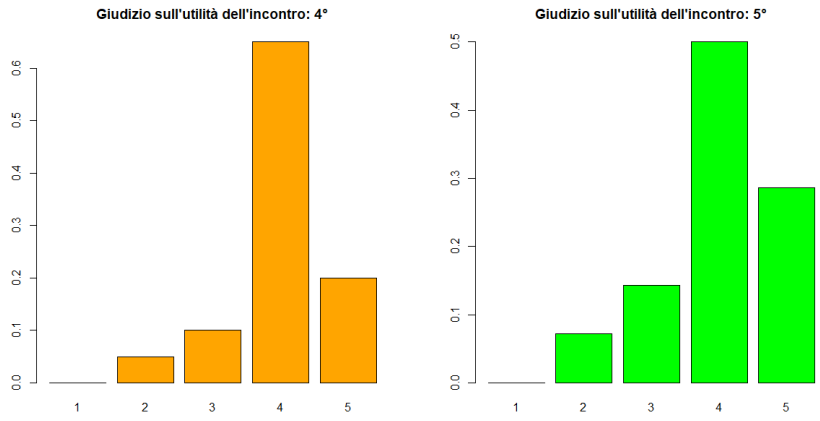


Figura 5.13: Giudizio sull'utilità: 4° vs 5°.

	Moda$_{IV}$	**Moda**$_{V}$	**Media**$_{IV}$	**Media**$_{V}$
Utilità	4	4	4.00	3.84

Alla base di ciò potrebbe esserci l'interesse scaturito da argomenti nuovi che hanno colto un'attenzione maggiore, nonostante ci si aspettasse una reazione in misura più grande da parte degli studenti del quinto anno. Infatti si pensava che degli esempi concreti di utilizzo di strumenti già conosciuti per la risoluzione problemi reali potessero interessare maggiormente gli studenti dell'ultimo anno. Il test basato sui ranghi rigetta l'ipotesi che ci siano differenze significative tra i due gruppi:

Variabile	**Statistica W**	**p-value**
Utilità	211	0.52

A conclusione di tutte le analisi condotte, si può affermare che l'intervento è stato efficiente nel suo complesso. Ha fornito un'idea chiara di come la Statistica svolga un ruolo fondamentale nella risoluzione di problemi quotidiani. L'incontro è servito agli insegnanti come stimolo per porre maggiore attenzione alla Statistica in classe, ma allo stesso tempo ha dato nuova luce a questa materia agli occhi dei ragazzi. Quest'esperienza ha saputo cogliere l'attenzione e l'interesse degli studenti tant'è che hanno fornito alcuni suggerimenti per progetti futuri. In generale i ragazzi consigliano di estendere questo tipo di esperienza anche ad altri settori (ad esempio il settore Informatico e non solo) per dimostrare l'interdisciplinarità della Statistica, con esempi reali e innovativi. Inoltre richiedono delle attività laboratoriali per maneggiare autonomamente i dati e condurre analisi simili a quelle riportate nell'incontro, nonchè la possibilità di svolgere più incontri per avere maggiori approfondimenti in merito a tematiche legate alla Statistica.
Tutto ciò a conferma di quanto sia utile e importante la divulgazione della Statistica nelle scuole, per riuscire ad indirizzare studenti e insegnanti in un percorso che preveda l'acquisizione delle conoscenze relative a questa materia.

Conclusione

In questa tesi si è cercato di sottolineare l'importanza dell'insegnamento della Statistica nelle scuole secondarie di secondo grado. La responsabilità di far acquisire le competenze di tale disciplina spetta agli insegnanti di Matematica. Infatti con la riforma emanata dal Ministero dell'Istruzione (MIUR) l'insegnamento della Statistica è stato accorpato a quello di Matematica costringendo, limitatamente alle loro competenze, i docenti di Matematica ad impartire lezioni inerenti a questa materia. Tuttavia la loro impronta puramente matematica non giova all'insegnamento della Statistica e tanto meno agli studenti. Perciò sono stati proposti due moduli didattici trattanti concetti statistici di base con un duplice scopo: agevolare i docenti ad insegnare una materia non propriamente di competenza; avvicinare i ragazzi ad una disciplina per molti sconosciuta, in modo che ne comprendano l'importanza non solo a livello scolastico.
Il primo modulo tratta i temi relativi a correlazione e regressione, ponendo particolare attenzione a concetti di Statistica di base (variabili e grafici). Nello specifico viene utilizzato un dataset che gli studenti hanno analizzato in un altro contesto disciplinare (Chimica).
Il secondo modulo ha un'impronta più innovativa e tratta la *Sentiment Analysis* applicata a Twitter. Lo scopo è intrigare gli studenti con una tematica molto vicina al loro quotidiano per renderli consapevoli di come la Statistica trovi applicazioni in svariati ambiti, anche innovativi.
Infine questi moduli sono stati proposti a due classi dell'indirizzo chimico dell'*I.T.I.S. E. Fermi* di Bassano del Grappa, grazie alla collaborazione di alcuni docenti di Matematica e Chimica. Durante l'incontro sono stati illustrati, in maniera sintetica, i due moduli didattici affrontati in questa tesi.

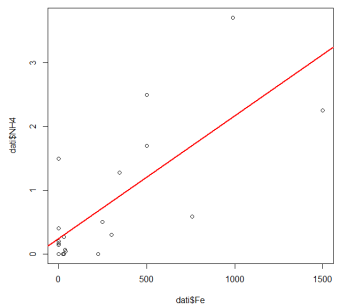
Inoltre sono stati presentati i corsi di Laurea in Statistica, nell'eventualità che qualche studente voglia iscriversi una volta terminato il percorso di studi. Alla fine è stato sottoposto ai ragazzi un questionario di gradimento per comprendere se, effettivamente, lo scopo prefissato sia stato raggiunto. Nel complesso gli studenti sono stati soddisfatti dall'incontro e i docenti positivi ad altre collaborazioni di questo tipo. La divulgazione della Statistica nelle scuole ha un ruolo molto importante per aiutare i docenti di Matematica ad insegnarla e a comprendere la sua importanza, in modo che vi diano più peso in classe. Soprattutto, però, divulgare in modo intelligente e interessante la Statistica aiuta i ragazzi ad aprire la mente verso una disciplina utile che risolve problemi concreti legati alla vita reale in diversi contesti.

Appendice A

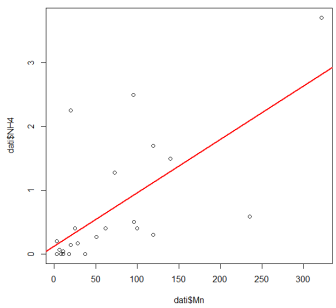
Codice R per il Monitoraggio delle acque potabili del Veneto

```
CARICO I DATI
> dati=read.csv2("acque_potabili.csv", header=T, dec=".")
> head(dati)

TRASFORMO LE VARIABILI IN NUMERICHE
> dati$NH4=as.numeric(dati$NH4) # Ammonio
> dati$Fe=as.numeric(dati$Fe) # Ferro
> dati$As=as.numeric(dati$As) # Arsenico
> dati$Mn=as.numeric(dati$Mn) # Manganese

MATRICE DI CORRELAZIONE
> cor(dati)
         As       NH4        Fe        Mn
As  1.0000000 0.5457497 0.4290931 0.1282860
NH4 0.5457497 1.0000000 0.7542977 0.6772078
Fe  0.4290931 0.7542977 1.0000000 0.5102653
Mn  0.1282860 0.6772078 0.5102653 1.0000000
```

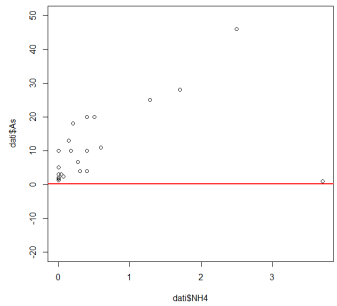
```
REGRESSIONE TRA FERRO E AMMONIO
> m=lm(dati$NH4~dati$Fe)
> summary(m)
Call:
lm(formula = dati$NH4 ~ dati$Fe)

Residuals:
Min      1Q  Median      3Q     Max
-1.1158 -0.2791 -0.1612  0.1598  1.5646

Coefficients:
            Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 0.2401460  0.1572061   1.528    0.141
dati$Fe     0.0019284  0.0003579   5.389    2.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6559 on 22 degrees of freedom
Multiple R-squared:  0.569,Adjusted R-squared:  0.5494
F-statistic: 29.04 on 1 and 22 DF,  p-value: 2.066e-05

> plot(dati$Fe,dati$NH4)
> abline(m, col="red", lwd=2)
```



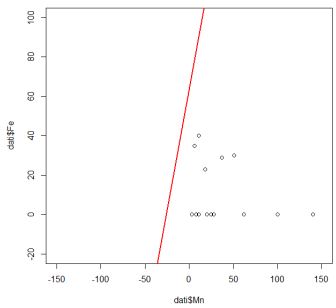
```
REGRESSIONE TRA AMMONIO E MANGANESE
> m1=lm(dati$NH4~dati$Mn)
> summary(m1)
Call:
lm(formula = dati$NH4 ~ dati$Mn)

Residuals:
Min      1Q  Median      3Q     Max
-1.5031 -0.2727 -0.1786  0.1038  1.9622

Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
(Intercept) 0.119836   0.199004   0.602   0.553213
dati$Mn     0.008397   0.001945   4.317   0.000278 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7351 on 22 degrees of freedom
Multiple R-squared:  0.4586,Adjusted R-squared:  0.434
F-statistic: 18.64 on 1 and 22 DF,  p-value: 0.0002782

> plot(dati$Mn,dati$NH4)
> abline(m1, col="red", lwd=2)
```



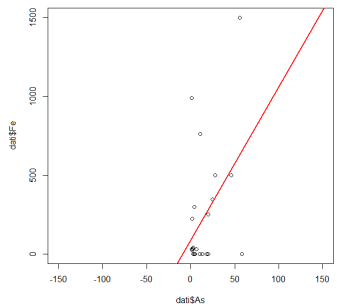
```
REGRESSIONE TRA AMMONIO E ARSENICO
> m2=lm(dati$NH4~dati$As)
> summary(m2)
Call:
lm(formula = dati$NH4 ~ dati$As)

Residuals:
Min      1Q  Median      3Q     Max
-0.5809 -0.3601 -0.2562  0.0390  3.4680

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  0.21026    0.23079    0.911   0.37216
dati$As      0.03170    0.01038    3.055   0.00581 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8371 on 22 degrees of freedom
Multiple R-squared:  0.2978,Adjusted R-squared:  0.2659
F-statistic: 9.332 on 1 and 22 DF,  p-value: 0.005806

> plot(dati$NH4,dati$As,ylim=c(-20,50))
> abline(m2, col="red",lwd=2)
```



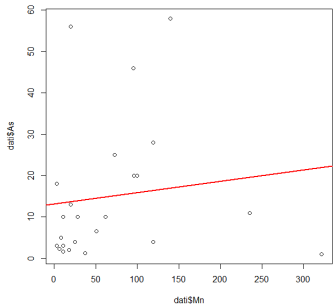
```
REGRESSIONE TRA FERRO E MANGANESE
> m3=lm(dati$Fe~dati$Mn)
> summary(m3)
Call:
lm(formula = dati$Fe ~ dati$Mn)

Residuals:
Min      1Q  Median      3Q     Max
-410.39 -125.97  -71.35  104.69 1386.58

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)  63.9289    90.9815    0.703   0.4896
dati$Mn       2.4747     0.8892    2.783   0.0108 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 336.1 on 22 degrees of freedom
Multiple R-squared:  0.2604,Adjusted R-squared:  0.2268
F-statistic: 7.745 on 1 and 22 DF,  p-value: 0.01085

> plot(dati$Mn,dati$Fe,ylim=c(-20,100),xlim=c(-150,150))
> abline(m3, col="red",lwd=2)
```



```
REGRESSIONE TRA FERRO E ARSENICO
> m4=lm(dati$Fe~dati$As)
> summary(m4)
Call:
lm(formula = dati$Fe ~ dati$As)

Residuals:
Min      1Q  Median      3Q     Max
-649.93 -181.99  -77.38   44.31  893.75

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)   84.506     97.310     0.868   0.3945
dati$As        9.749      4.375     2.228   0.0364 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353 on 22 degrees of freedom
Multiple R-squared:  0.1841,Adjusted R-squared:  0.147
F-statistic: 4.965 on 1 and 22 DF,  p-value: 0.03641

> plot(dati$As,dati$Fe, xlim=c(-150,150))
> abline(m4, col="red",lwd=2)
```



```
REGRESSIONE TRA ARSENICO E MANGANESE
> m5=lm(dati$As~dati$Mn)
> summary(m5)
Call:
lm(formula = dati$As ~ dati$Mn)

Residuals:
Min      1Q  Median      3Q     Max
-20.927 -10.552  -6.357   4.398  42.343

Coefficients:
            Estimate    Std. Error t value Pr(>|t|)
(Intercept) 13.10951    4.61790    2.839   0.00955 **
dati$Mn      0.02738    0.04514    0.607   0.55024
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.06 on 22 degrees of freedom
Multiple R-squared:  0.01646,Adjusted R-squared:  -0.02825
F-statistic: 0.3681 on 1 and 22 DF,  p-value: 0.5502

> plot(dati$Mn,dati$As)
> abline(m5, col="red",lwd=2)
```

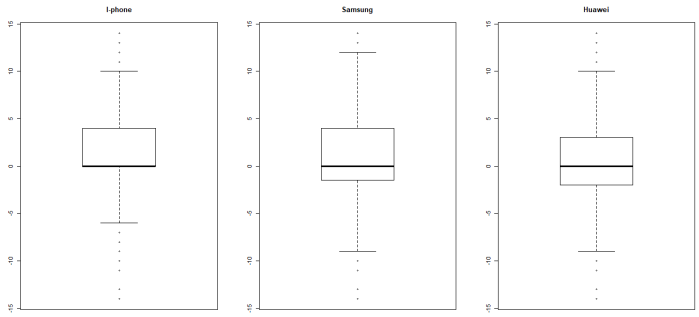


Appendice B

Codice R per la Sentiment Analysis

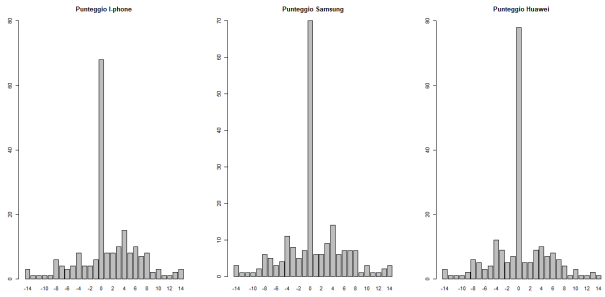
```
CARICO I DATI
> ip=read.csv2("iphone.csv",header=T)
> sam=read.csv2("samsung.csv",header=T)
> hu=read.csv2("huawei.csv",header=T)


ANALISI DESCRITTIVE DELLA VARIABILE SOMMA

Trasformo le variabili in tipo numerico
> ip$somma=as.numeric(ip$somma)
> sam$somma=as.numeric(sam$somma)
> hu$somma=as.numeric(hu$somma)

Calcolo delle medie
> mean(ip$somma)
> mean(sam$somma)
> mean(hu$somma)
```

```
Calcolo delle deviazioni standard
> sd(ip$somma)
> sd(sam$somma)
> sd(hu$somma)

Calcolo degli scarti interquartili
> IQR(ip$somma)
> IQR(sam$somma)
> IQR(hu$somma)

Boxplot
> par(mfrow=c(1,3))
> boxplot(ip$somma)
> boxplot(sam$somma)
> boxplot(hu$somma)
```




```
TRASFORMO LE VARIABILI IN TIPO FACTOR
> ip$score=as.factor(ip$score)
> sam$score=as.factor(sam$score)
> hu$score=as.factor(hu$score)
> ip$somma=as.factor(ip$somma)
> sam$somma=as.factor(sam$somma)
> hu$somma=as.factor(hu$somma)
```

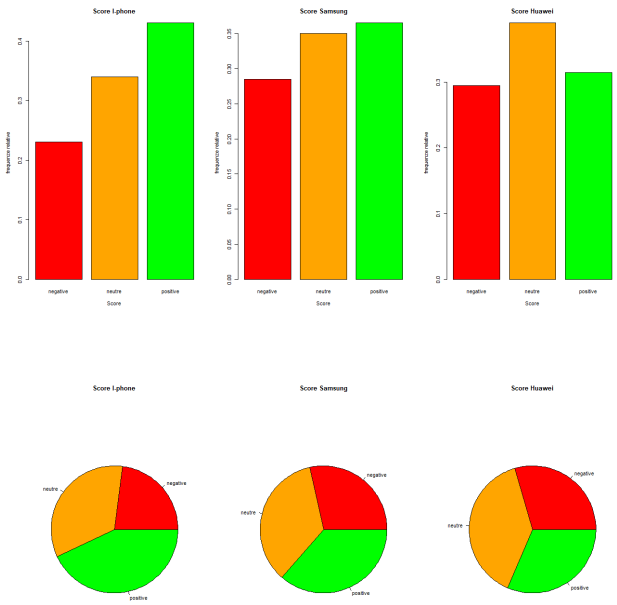
```
GRAFICI DEL PUNTEGGIO
> par(mfrow=c(1,3))
> plot(ip$somma,col="grey",main="Punteggio I-phone",ylim=c(0,80))
> plot(sam$somma,col="grey",main="Punteggio Samsung")
> plot(hu$somma,col="grey",main="Punteggio Huawei",ylim=c(0,80))
```




```
> n=200

GRAFICI SCORE I-PHONE
> par(mfrow=c(1,3))
> score_neg=nrow(ip[ip$score==-1,])/n
> score_neu=nrow(ip[ip$score==0,])/n
> score_pos=nrow(ip[ip$score==1,])/n
> barplot(c(score_neg,score_neu,score_pos),names.arg = c("negative",
+         "neutre","positive"),xlab="Score",ylab="frequenze
+          relative",main="Score I-phone",col=c("red","orange","green"))
> pie(c(score_neg,score_neu,score_pos),labels = c("negative","neutre",
+   "positive"),main="Score I-phone",col=c("red","orange","green"))


GRAFICI SCORE SAMSUNG
> score_neg=nrow(sam[sam$score==-1,])/n
> score_neu=nrow(sam[sam$score==0,])/n
> score_pos=nrow(sam[sam$score==1,])/n
```

```
> barplot(c(score_neg,score_neu,score_pos),names.arg = c("negative",
+         "neutre","positive"),xlab="Score",ylab="frequenze
+         relative",main="Score Samsung",col=c("red","orange","green"))
> pie(c(score_neg,score_neu,score_pos),labels = c("negative","neutre",
+     "positive"),main="Score Samsung",col=c("red","orange","green"))
```

```
GRAFICI SCORE HUAWEI
> score_neg=nrow(hu[hu$score==-1,])/n
> score_neu=nrow(hu[hu$score==0,])/n
> score_pos=nrow(hu[hu$score==1,])/n
> barplot(c(score_neg,score_neu,score_pos),names.arg = c("negative",
+         "neutre","positive"),xlab="Score",ylab="frequenze
+         relative",main="Score Huawei",col=c("red","orange","green"))
> pie(c(score_neg,score_neu,score_pos),labels = c("negative","neutre",
+     "positive"),main="Score Huawei",col=c("red","orange","green"))
```



Bibliografia e sitografia

Bibliografia

- Agresti, A., Finlay, B. (2009). *Statistica per le Scienze Sociali.* Pearson.
- Agresti, A., Franklin, C. (2013). *Statistica: l'arte e la scienza d'imparare dai dati.* Pearson.
- Baroncini, P., Manfredi, R. (2016). *MultiMath - Dati e previsioni.* Ghisetti e Corvi.
- Bergamini, M., Barozzi, G., Trifone, A. (2016). *Matematica.blu.* Zanichelli.
- Bernstein, S., Bernstein, R. (2003). *Statistica Descrittiva.* McGraw-Hill.
- Ceron, A., Currini, L., Iacus, S. (2013). *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete.* Springer.
- Cicchitelli, G. (2012). *Statistica: principi e metodi.* Pearson.
- Diamond, I., Jefferies, J. (2002). *Introduzione alla statistica. Per le scienze sociali.* McGraw-Hill.
- Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* M&C.
- Pace, L., Salvan, A. (1996). *Introduzione alla Statistica. I Statistica descrittiva.* Cedam.
- Piccola, D. (2010). *Statistica.* il Mulino.

- Rosenthal, J.S. (2005). *Le Regole del Caso: Istruzioni per l'Uso*. Longanesi.
- Sasso, L. (2012). *La matematica a colori - edizione verde*. Petrini.
- Spiegel, M. R. (2003). *Probabilità e Statistica*. Schaum.
- Trovato, M. (2004). *Probabilità, Statistica e ricerca operativa. Metodi e strumenti*. Ghisetti e Corvi editori.

Sitografia

- `www.amstat.org`;
- `www.arpa.veneto.it`;
- `www.disia.unifi.it`;
- `www.istat.it`;
- `www.istruzione.it`;
- `www.miur.gov.it`;
- `www.orizzontescuola.it`;
- `www.pls.scienze.unipd.it`;
- `www.r-project.org`;
- `www.sis-statistica.it`;
- `www.scuolainforma.it`;
- `www.tfa.cineca.it`;
- `www.tuttoscuola.com`;
- `www.twitter.com`.