Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea in Matematica

UN'ESPOSIZIONE IPERTESTUALE DI ALCUNI ELEMENTI DI STATISTICA DESCRITTIVA

Tesi di Laurea in Statistica Matematica

Relatore:
Chiar.ma Prof.
EMANUELA CALICETI

Correlatore:
Chiar.mo Prof.
ALESSANDRO GIMIGLIANO

Presentata da:
MIRIAM NIERI

III Sessione
Anno Accademico 2011-2012

Che soddisfazione si può provare
a non capire qualcosa?
Raymond Queneau, "Odile", 1937.

Introduzione

Questa tesi si inserisce nell'ambito del Progetto Matematic@ curato dai professori D. Aliffi e A. Gimigliano del Dipartimento di Matematica dell'Università di Bologna. Tale progetto si pone l'obiettivo di presentare alcuni argomenti di base di matematica, in particolare quelli del primo biennio delle facoltà scientifiche, in maniera multimediale, allo scopo di renderli più accattivanti, più accessibili e facilmente fruibili agli studenti che intraprendono un percorso universitario.
In particolare questa tesi si pone l'obiettivo di realizzare un'esposizione ipertestuale di alcuni elementi di statistica descrittiva, cercando di integrare efficacemente gli aspetti teorici e quelli applicativi. Nello specifico questi ultimi sono basati su esempi illustrati in dettaglio e su autovalutazioni della preparazione proposte all'utente in maniera interattiva attraverso test a risposta multipla ed esercizi guidati. Grazie alla struttura ipertestuale il lettore è inoltre libero di scegliere un proprio percorso all'interno dei vari contenuti messi a disposizione.
Questa tesi si struttura in due capitoli. Il Capitolo 1 inizialmente si interroga sulle cause delle difficoltà nell'apprendimento della matematica da parte di un elevato numero di studenti, facendo riferimento agli articoli di Zan R. [8] e Berengo F. [1]; dopodiché viene illustrato il metodo dell'e-learning e la sua utilità nella didattica utilizzando l'articolo di Di Martino P., Fiorentino G. e Zan R. [4], unitamente alle fonti [6] e [7]. Infine viene svolta una accurata descrizione della struttura e delle funzionalità dell'ipertesto realizzato. Il Capitolo 2 presenta le nozioni di base della statistica descrittiva sia mono-

variata che bivariata contenute nell'ipertesto, seguendo una linea espositiva tradizionale. Ispirandosi ai testi di Di Ciaccio A. e Borra S. [2, 3], oltre che al trattato [5], a partire dall'introduzione della dovuta terminologia per lo studio dei fenomeni collettivi, vengono innanzitutto introdotti i primi strumenti per l'analisi dei dati e per una rappresentazione sintetica dei risultati di una rilevazione statistica, con particolare riguardo ai caratteri quantitativi. Vengono dunque introdotte ed illustrate le principali medie analitiche e i più usati indici di variabilità. Nei paragrafi successivi l'attenzione si focalizza sull'analisi dell'associazione tra due caratteri quantitativi: dall'indipendenza statistica alla dipendenza perfetta. Infine per completare il quadro della statistica bivariata, viene illustrato il modello di regressione lineare semplice, attraverso una presentazione accurata del metodo di interpolazione dei minimi quadrati per l'individuazione della retta di regressione.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Apprendimento della matematica e multimedialità

Il passaggio dalla scuola superiore all'università risulta essere faticoso per molti studenti, tanto da poter essere considerato una delle cause della cosiddetta *mortalità universitaria*. In particolare, nei corsi di laurea delle facoltà scientifiche assumono un ruolo molto importante le difficoltà riscontrate nei corsi di matematica, tanto che si potrebbe persino arrivare a parlare di *mortalità matematica*, come fa Zan R. in [8]. Sorgono spontanee varie domande: perchè? Quali sono le cause di tanti fallimenti ed abbandoni? Da cosa è causato questo grande disagio nei confronti della matematica? E cosa si può fare per affrontare tali problemi?

1.1 Difficoltà nell'apprendimento della matematica

I motivi delle difficoltà nell'acquisizione dei concetti matematici sono molteplici. Innanzitutto la matematica, fra tutte le discipline scientifiche, è unanimemente riconosciuta come quella in cui la sequenzialità è più importante: dopo la scuola elementare ogni conoscenza ed abilità si fonda su conoscenze e abilità già acquisite in precedenza. Vanno inoltre considerate le difficoltà

legate all'utilizzo del linguaggio formale, dell'impianto assiomatico-deduttivo e del rigore logico richiesto nello sviluppo dei ragionamenti. Di determinante importanza sono poi le componenti di natura metacognitiva, motivazionale ed affettiva. Insorgono infatti convinzioni (su di sé, sul successo in matematica, sulle aspettative dei docenti) che contribuiscono alla nascita e al consolidamento di atteggiamenti negativi nei confronti della disciplina, dovuti invece in gran parte ad uno scarso sviluppo di consapevolezza di sé e delle proprie risorse. Tra gli atteggiamenti negativi annoveriamo certamente la noia e il non interesse per qualcosa di cui non si percepisce il senso: non viene data importanza al fatto che le richieste possano avere un senso, né tanto meno al fatto che lo possano avere i risultati. Tutto il processo logico-deduttivo viene ridotto ad un mero esercizio di calcolo e ad un'estenuata ricerca di un risultato, qualunque esso sia.
L'obiettivo prioritario dell'educazione scolastica attuale è invece tutt'altro. Oggi più che mai, in una società caratterizzata da un'enorme quantità di informazioni in costante evoluzione, in cui la dotazione culturale del singolo è sempre meno statica e sempre più dipendente dall'incessante ritmo con cui mutano le fonti informative, ciò che è basilare è il "saper imparare". Diventa allora fondamentale, piuttosto che avere un corpo stabile di conoscenze e informazioni, saper sviluppare la capacità di acquisirle in modo autonomo e continuativo. Quanto detto ha valenza in ogni contesto disciplinare, ma chiaramente esso assume rilevanza particolare per quel che riguarda la matematica, in quanto in tale disciplina la conoscenza e la gestione dei risultati deve poggiare sempre sulla consapevolezza dei processi sottostanti.

1.2 Utilità del metodo dell'e-learning

Si percepisce dunque la necessità di coinvolgere maggiormente lo studente nel processo formativo, facendolo agire in prima persona, richiedendogli un impegno personale nella costruzione dei concetti. Durante la didattica classica, ossia in presenza, è quasi impossibile poter fornire ad ogni studente

il cosiddetto approccio "per scoperta" (come lo chiama Berengo F. in [1]), che gli consenta di venire in contatto con una nuova informazione in modo attivo e totalmente autonomo. L'approccio di tipo informatico invece, attraverso l'utilizzo della multimedialità, oltre a risultare gradevole e di per sé motivante per la maggior parte degli studenti, può persino attivare processi di apprendimento che coinvolgono stili cognitivi diversi.
Per apprendimento on-line, cosiddetto e-learning, viene inteso "l'uso delle tecnologie multimediali e di Internet per migliorare la qualità dell'apprendimento facilitando l'accesso alle risorse e ai servizi così come anche agli scambi in remoto e alla collaborazione (creazione di comunità virtuali di apprendimento)". Innanzitutto risulta evidente come tale tipo di apprendimento consenta la dilatazione del tempo e dello spazio: esso infatti sfrutta le potenzialità rese disponibili da Internet per fornire al discente contenuti ai quali poter accedere in qualsiasi momento e in ogni luogo esista una connessione Internet.
In particolare l'e-learning possiede tre caratteristiche fondamentali che lo differenziano dalla didattica tradizionale: l'interattività, ossia il coinvolgimento dello studente attraverso il *learning by doing*; la dinamicità, richiesta allo studente stesso, di acquisire competenze mirate *just in time*; la modularità dei contenuti, organizzata secondo gli obiettivi formativi preposti.
Il vero punto di forza dell'e-learning sta dunque nella possibilità data al discente di personalizzare il proprio percorso, scegliendo di volta in volta se e quando utilizzare una determinata risorsa. Lo studente viene quindi reso protagonista attivo del proprio percorso, responsabilizzato delle varie decisioni e della valutazione della propria preparazione. Inoltre, attraverso la rete, egli può accedere a materiali prodotti da persone diverse, secondo ottiche diverse, avendo quindi la possibilità di scegliere ciò che trova più utile e comprensibile sulla base della conoscenza che ha di sé e della propria preparazione.
Si noti che l'e-learning acquista grande sostenibilità didattico-formativa nel momento in cui diventa un reale valore aggiunto alla didattica, sfruttando efficientemente le potenzialità delle tecnologie nel processo educativo dello

studente in modo da arricchirlo e migliorarlo, raggiungendo obiettivi che non potrebbero essere raggiunti con strumenti ed approcci tradizionali.

1.3 Progetto Matematic@ e realizzazione di un'esposizione ipertestuale

In quest'ottica di personalizzazione attiva dell'apprendimento e autovalutazione della preparazione si inserisce il Progetto Matematic@ curato dai professori Alessandro Gimigliano e Davide Aliffi del Dipartimento di Matematica dell'Alma Mater Studiorum di Bologna. Tale progetto si pone un duplice scopo: da una parte quello di presentare la matematica delle facoltà scientifiche, in particolare quella del primo biennio, in maniera multimediale, dall'altra quello della divulgazione della matematica. Il Progetto Matematic@ si sviluppa in un sito consultabile all'indirizzo *http://progettomatematica.dm.unibo.it/* la cui costruzione procede principalmente attraverso il lavoro di studenti del Corso di Laurea in Matematica mediante l'assegnazione di attività di tirocinio o di Tesi di Laurea. Ed è proprio quest'ultimo il caso che mi riguarda personalmente: il lavoro di questa tesi è stato quello di realizzare un ipertesto che contenesse elementi di statistica descrittiva, presentati in maniera interattiva, che potesse entrare a far parte del Progetto Matematic@ sopracitato.

Per la realizzazione di questo ipertesto è stato utilizzato *DreamweaverMX 2004*, programma per la realizzazione di siti web prodotto da *Macromedia*. Tale programma ha un'interfaccia grafica molto intuitiva del tipo WYSIWYG (*What You See Is What You Get*) e consente di affiancare la programmazione in linguaggio HTML con la visualizzazione immediata del prodotto finale.
La Figura 1.1 riporta la Home page del sito realizzato. Come si può vedere essa presenta un breve indice che consente di accedere a tutte le principali funzionalità del sito. Viene infatti fornita innanzitutto un'introduzione agli argomenti di statistica descrittiva che verranno presentati, unitamente ad una breve guida alla navigazione che illustra come muoversi all'interno del

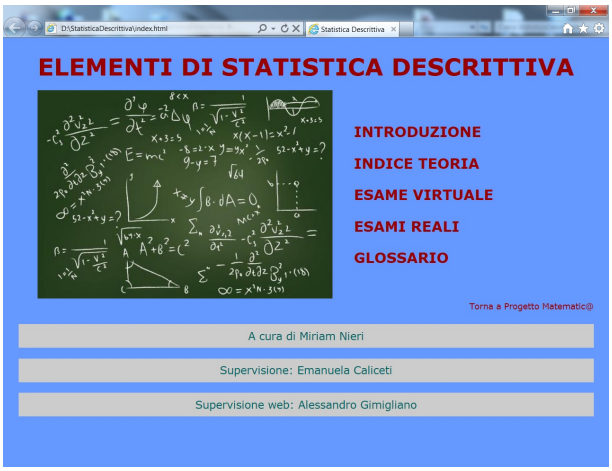

Figura 1.1: Home page

sito. Inoltre scegliendo la voce "Torna a Progetto Matematic@" viene evidentemente data la possibilità di tornare al sito del Progetto Matematic@, per poter consultare ipertesti relativi ad altri argomenti.

La voce "Indice teoria" consente di passare a un'altra pagina, mostrata in

Figura 1.2: Indice capitoli teorici

Figura 1.2, dove sono elencati i titoli dei dieci capitoli in cui è stata suddivisa la presentazione degli elementi di statistica descrittiva. Da questo indice è

possibile accedere alla sezione che più si desidera consultare, oppure si può tornare alla Home page mediante il bottone Home raffigurante una casetta stilizzata sulla quale compare la parola "home" al passaggio del mouse, opzione realizzata tramite la funzione Rollover Image.

Cliccando sulla voce "Glossario" si visualizza la pagina riportata in Figura

Figura 1.3: Glossario dei contenuti

1.3. Qui sono riportati in ordine alfabetico i termini più importanti della trattazione teorica degli elementi di statistica descrittiva; cliccando su ognuno di essi si viene subito rimandati all'esatto punto della presentazione in cui essi compaiono. Questo glossario dà quindi la possibilità all'utente, nella maniera più assoluta, di personalizzare il suo percorso all'interno dei contenuti resi disponibili.

Una sezione molto importante è quella relativa agli "Esami reali", pagina riportata in Figura 1.4. Qui sono forniti vari testi d'esame, divisi per anno, mediante i quali l'utente è invitato a misurare la propria preparazione. In particolare sono anche fornite varie soluzioni, affinché egli possa poi verificare l'effettiva correttezza degli esercizi svolti. L'aspetto più interessante di questa pagina sono però gli esercizi guidati: tali esercizi focalizzano l'attenzione sul corpo centrale dell'esposizione teorica, ossia sull'analisi dell'associazione tra due caratteri quantitativi; essi guidano il lettore dallo studio della tabella

Figura 1.4: Esercizi proposti

a doppia entrata e del grafico di dispersione, passando per importanti indici come la covarianza e il coefficiente di correlazione lineare, fino al metodo di interpolazione dei minimi quadrati, con lo studio dei coefficienti di regressione e dell'indice di determinazione.

Alla voce "Esame virtuale" ciò che viene proposto è un questionario com-

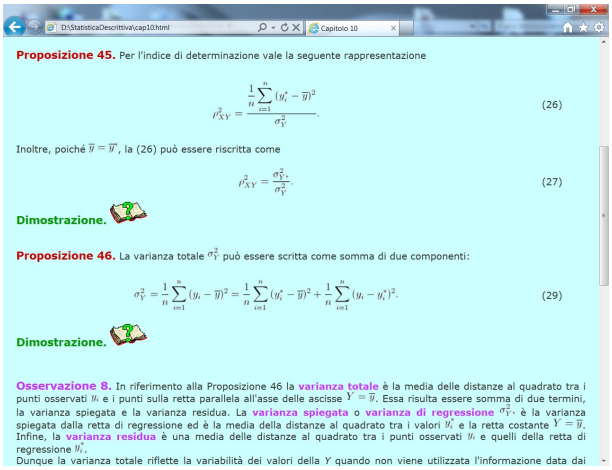

Figura 1.5: Esempio di capitolo della sezione teorica

posto da dieci domande a risposta multipla del valore di tre punti ciascuna,

che l'utente è invitato a compilare. Per ogni domanda, alla convalida della risposta scelta, compare una piccola finestrella in cui è riportato l'esito della risposta data; nel caso di responso negativo viene fornita inoltre una breve delucidazione. Al termine della compilazione del questionario, cliccando su "Calcola il totale" viene dato all'utente un punteggio in trentesimi che gli consentirà di ricevere una prima valutazione della sua personale preparazione relativamente agli argomenti trattati e che lo indirizzerà quindi verso un eventuale ripasso mirato delle lacune presentate.

Ora analizziamo meglio, sempre dal punto di vista ipertestuale, la sezione

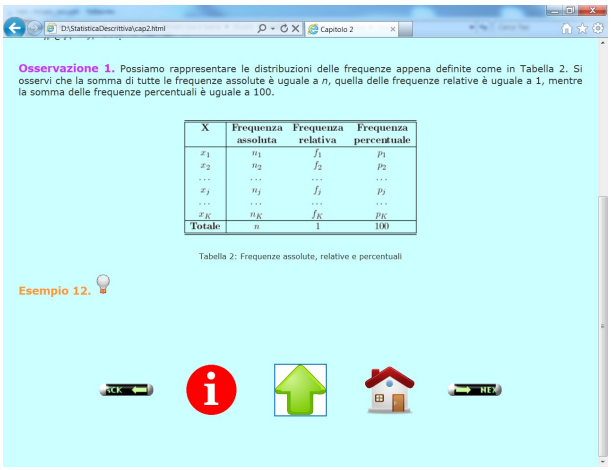

Figura 1.6: Icone poste al termine di ogni capitolo

dedicata all'esposizione teorica degli elementi di statistica descrittiva. Si ricorda infatti che i contenuti presenti in tale sezione saranno approfonditi in dettaglio nel Capitolo 2 di questa tesi.

È da notare innanzitutto che le formule e le tabelle presenti nell'ipertesto realizzato, scritte in precedenza in formato TEX, sono state tutte trasformate in immagini in formato GIF, utilizzabili quindi dal programma *DreamweaverMX*, grazie alla pagina $http://www.codecogs.com/latex/eqneditor.php$. Ognuno dei dieci capitoli teorici ha un'impostazione simile a quella riportata in Figura 1.5. Come si può osservare viene innanzitutto fatto ampio utilizzo dei colori per sottolineare e allo stesso tempo differenziare definizioni, propo-

sizioni, osservazioni, dimostrazioni, ecc. Al termine di ogni pagina è riportata una serie di icone, come si può vedere in Figura 1.6, con diverse funzioni: le due più esterne consentono di passare rapidamente da un capitolo teorico al suo precedente o al suo successivo (mediante collegamenti ipertestuali creati sulle immagini GIF); le icone più interne invece, che grazie alla funzione Rollover Image cambiano immagine al passaggio del muose, consentono di tornare all'indice dei capitoli teorici, all'inizio del capitolo stesso (cioè a inizio pagina) oppure di rientrare alla Home page del sito.

Si noti inoltre che, allo scopo di alleggerire la lettura di ogni singolo capi-

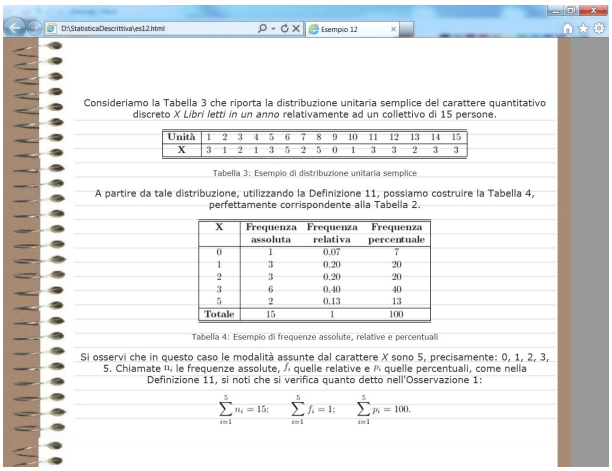

Figura 1.7: Esempio di esercizio della sezione teorica

tolo, gli esempi e le dimostrazioni non sono riportati nel corpo centrale della pagina teorica, ma sono visualizzabili solo tramite click del mouse sulle relative icone (immagini in formato GIF), azione che aprirà una nuova pagina del browser ad essi dedicata. Per differenziare queste pagine si è scelto di utilizzare sfondi differenti, come si può osservare in Figura 1.7.

In conclusione con questo sito si è cercato di presentare alcuni elementi di statistica descrittiva in maniera multimediale e interattiva, allo scopo di rendere tale argomento più accattivante, invitante, facilmente fruibile e di più semplice comprensione, anche grazie ai numerosi esempi ed esercizi forniti per l'autovalutazione. In questo sito ogni fruitore viene reso responsabile

della scelta di un percorso di apprendimento personalizzato e maggiormente consapevole del livello di preparazione raggiunto sull'argomento.

Capitolo 2

Elementi di statistica descrittiva

In questo capitolo verranno introdotti ed illustrati alcuni elementi di base di statistica descrittiva sia monovariata che bivariata. La presentazione di questi argomenti si basa principalmente sui testi di Di Ciaccio A. e Borra S. [2, 3] sull'argomento, in particolare vista la funzionalità dell'impostazione didattica utilizzata dai due autori. Anche nella presente esposizione infatti si è cercato di affiancare sempre le nozioni date con esempi esplicativi per una migliore e più approfondita comprensione da parte del lettore, utilizzando a questo scopo anche il trattato [5] di Mecatti F.

Il capitolo ha inizio con l'introduzione, nel Paragrafo 1, della terminologia statistica di base per lo studio dei fenomeni collettivi. Segue poi la presentazione di strumenti idonei ad una rappresentazione sintetica dei dati e dei risultati di una rilevazione statistica: dalla distribuzione di frequenze nel Paragrafo 2, a specifici indici in grado di evidenziare le caratteristiche essenziali della distribuzione di un carattere, come le medie analitiche nel Paragrafo 3 e gli indici di variabilità nel Paragrafo 4. Nel Paragrafo 5 si procede con l'analisi congiunta di due caratteri mediante la tabella a doppia entrata.

Dal Paragrafo 6 ha inizio il corpo centrale dell'esposizione: lo studio dell'associazione tra due caratteri, dall'indipendenza statistica alla dipendenza

perfetta. Dal Paragrafo 7 l'analisi si concentra in particolar modo sullo studio dell'associazione tra caratteri quantitativi, introducendo indici come la covarianza e il coefficiente di correlazione lineare. Nel Paragrafo 8 viene poi illustrato il modello di regressione lineare semplice per l'individuazione di una funzione che possa descrivere sinteticamente il legame che unisce due caratteri quantitativi; in particolare viene studiato in dettaglio il metodo di interpolazione dei minimi quadrati per l'individuazione della retta di regressione nel Paragrafo 9. A partire dai risultati di quest'ultimo, il Paragrafo 10 considera infine l'indice di determinazione in termini di varianza totale, spiegata e residua.
Per evidenti ragioni di brevità la maggior parte dei risultati è solo enunciata, senza dimostrazione. Per ulteriori dettagli ed approfondimenti il lettore può riferirsi ai già citati trattati [2, 3], oltre che a [5].

2.1 Collettivi statistici, caratteri, modalità

La statistica analizza in termini quantitativi i fenomeni collettivi. La Tabella 2.1 riporta un piccolo insieme di osservazioni. Ogni riga della tabella

Nome	**Età**	**Sesso**	**Titolo di studio**	**Attività**
Bianchi	26	M	diploma	studente
Ferrari	50	F	diploma	casalinga
Neri	46	M	laurea	disoccupato
Rossi	32	M	laurea	occupato
Villa	39	F	laurea	occupato

Tabella 2.1: Esempio di tabella di dati

corrisponde a un individuo del quale sono rilevati *Nome*, *Età*, *Sesso*, *Titolo di Studio* e *Attività*. Ciascuna di queste caratteristiche o più brevemente **caratteri**, assume in corrispondenza di ogni individuo una determinata **mo-**

dalità. Per esempio il carattere *Attività* assume la modalità *studente* in corrispondenza dell'individuo *Bianchi*.

Definizione 2.1. Si definisce **unità statistica** l'unità elementare su cui vengono osservati i caratteri oggetto di studio. Un insieme di unità statistiche omogenee rispetto ad una o più caratteristiche costituisce un **collettivo** statistico o una popolazione.

Nella Tabella 2.1 l'unità elementare è l'individuo, ma in generale potrebbe essere un oggetto, un territorio, un tempo, ecc.
Un carattere può assumere modalità differenti in corrispondenza delle diverse unità statistiche del collettivo. Ciò che è fondamentale è che le modalità siano sempre **esaustive**, cioè in grado di rappresentare tutti i possibili modi di essere di un carattere, e **non sovrapposte**, affinché ad ogni unità si possa associare una sola modalità.

Definizione 2.2. Quando le modalità sono espresse numericamente il carattere è detto **quantitativo** (o variabile), altrimenti è detto **qualitativo** (o mutabile).

Esempio 2.3. Facendo riferimento alla Tabella 2.1 si può notare come il carattere *Età* sia quantitativo, mentre i caratteri *Sesso*, *Titolo di Studio* e *Attività* siano di tipo qualitativo.

Definizione 2.4. Un carattere qualitativo può essere **sconnesso** (o su scala nominale), se date due sue modalità è possibile solo affermare se queste sono uguali o diverse, oppure **ordinato** (o su scala ordinale) se le sue modalità possono essere ordinate, cioè date due sue modalità è possibile specificare quale delle due precede l'altra.

Esempio 2.5. Caratteri sconnessi sono: *Sesso*, *Attività*, *Luogo di Nascita*, *Religione*. Caratteri ordinati sono: *Grado di soddisfazione* (con modalità Poco, Abbastanza, Molto) oppure *Titolo di studio* (con modalità Senza titolo, Licenza elementare, Licenza media, Diploma, Laurea, Dottorato).

Definizione 2.6. I caratteri quantitativi vengono distinti in discreti e continui. In un carattere quantitativo **discreto** l'insieme delle modalità assumibili può essere messo in corrispondenza biunivoca con un sottoinsieme dei numeri interi. In un carattere quantitativo **continuo** tale insieme può invece essere messo in corrispondenza biunivoca con un intervallo di $\mathbb{R}$ (insieme dei numeri reali).

Esempio 2.7. Caratteri quantitativi discreti sono: *Numero di figli*, *Numero di pezzi prodotti*, *Posto in graduatoria*. Un carattere continuo è invece il *Peso*.

Definizione 2.8. Se il carattere è quantitativo si definisce **suddivisione in classi** del carattere l'operazione consistente nel suddividere l'insieme dei possibili valori in intervalli tra loro disgiunti. È opportuno definire le classi in modo tale che:

1. il loro numero sia abbastanza piccolo da fornire una sintesi adeguata, ma sufficientemente grande da mantenere l'informazione con un livello accettabile di dettaglio;

2. siano tra loro disgiunte;

3. comprendano tutte le possibili modalità del carattere;

4. abbiano, se possibile, la stessa ampiezza.

Esempio 2.9. Consideriamo per esempio il carattere quantitativo continuo *Altezza*, misurata in centimetri. Nello stabilire gli estremi delle classi occorre tener presente che ognuna delle determinazioni del carattere deve essere compresa in una e una sola classe. Si rende allora necessario includere nella classe uno solo dei due estremi dell'intervallo, costruendo quindi intervalli del tipo $[140, 160[$, $[160, 180[$, $[180, 200[$ o viceversa fissando intervalli aperti a sinistra e chiusi a destra.

2.2 Distribuzioni unitarie e distribuzioni di frequenze

Definizione 2.10. Definiamo **distribuzione unitaria semplice** di un carattere l'elencazione delle modalità osservate, unità per unità, nel collettivo preso in esame. Si parla di **distribuzione unitaria multipla** quando tale elencazione si riferisce a più di un carattere.

La distribuzione unitaria, che per esempio è multipla nel caso della Tabella 2.1, pur descrivendo fedelmente la situazione osservata, non consente di cogliere in maniera sintetica le caratteristiche del fenomeno. Per ottenere una maggiore sintesi è necessario considerare la frequenza con cui le diverse modalità sono state osservate.

Definizione 2.11. Sia X un carattere con K modalità $x_1, \ldots, x_K$ osservato su un collettivo di n individui. $\forall i = 1, \ldots, K$ si definisce **frequenza assoluta** n_i della modalità x_i il numero di volte che tale modalità viene osservata nel collettivo. $\forall i = 1, \ldots, K$ si definiscono poi **frequenza relativa** e **frequenza percentuale** i-esime rispettivamente $f_i = \frac{n_i}{n}$ e $p_i = f_i \cdot 100$.

X	**Frequenza assoluta**	**Frequenza relativa**	**Frequenza percentuale**
x_1	n_1	f_1	p_1
x_2	n_2	f_2	p_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_j	n_j	f_j	p_j
$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_K	n_K	f_K	p_K
Totale	n	1	100

Tabella 2.2: Frequenze assolute, relative e percentuali

Osservazione 1. Possiamo rappresentare le distribuzioni delle frequenze appena definite come in Tabella 2.2. Si osservi che la somma di tutte le frequenze assolute è uguale a n, quella delle frequenze relative è uguale a 1, mentre la somma delle frequenze percentuali è uguale a 100.

Esempio 2.12. Consideriamo la Tabella 2.3 che riporta la distribuzione unitaria semplice del carattere quantitativo discreto X *Libri letti in un anno* relativamente ad un collettivo di 15 persone. A partire da tale distribuzione, utilizzando la Definizione 2.11, possiamo costruire la Tabella 2.4, perfettamente corrispondente alla Tabella 2.2. Si osservi che in questo caso le modalità assunte dal carattere X sono 5, precisamente: 0, 1, 2, 3, 5. Chiamate n_i le frequenze assolute, f_i quelle relative e p_i quelle percentuali, come nella Definizione 2.11, si noti che si verifica quanto detto nell'Osservazione 1:

$$\sum_{i=1}^{5} n_i = 15; \qquad \sum_{i=1}^{5} f_i = 1; \qquad \sum_{i=1}^{5} p_i = 100.$$

Unità	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	3	1	2	1	3	5	2	5	0	1	3	3	2	3	3

Tabella 2.3: Esempio di distribuzione unitaria semplice

2.3 Medie analitiche

Per descrivere l'insieme delle osservazioni di un carattere su di un collettivo possiamo impiegare una delle distribuzioni di frequenze appena viste, affidarci a opportune rappresentazioni grafiche, oppure possiamo limitarci a riportare il valore di precisi indici che evidenzino le caratteristiche essenziali della distribuzione del carattere. Un primo tipo di tali indici sono le **medie**. Esse possono essere **analitiche**, cioè calcolate attraverso operazioni

X	Frequenza assoluta	Frequenza relativa	Frequenza percentuale
0	1	0,07	7
1	3	0,20	20
2	3	0,20	20
3	6	0,40	40
5	2	0,13	13
Totale	15	1	100

Tabella 2.4: Esempio di frequenze assolute, relative e percentuali

algebriche sui valori del carattere, che dovrà essere perciò di tipo quantitativo, oppure di **posizione**, che possono essere determinate anche su caratteri di tipo qualitativo in quanto non utilizzano tali operazioni.

Definizione 2.13. Dato un insieme di n valori $x_1, \ldots, x_n$ di un carattere quantitativo X, la sua **media aritmetica** è definita da

$$M_a = M_a(X) = \overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j. \tag{2.1}$$

Se il carattere X è quantitativo discreto e conosciamo la sua distribuzione di frequenze, possiamo calcolare più velocemente la media aritmetica in questo modo:

$$M_a = \frac{1}{n} \sum_{j=1}^{K} x_j n_j \qquad \text{oppure} \qquad M_a = \sum_{j=1}^{K} x_j f_j \tag{2.2}$$

dove $x_1, \ldots, x_K$ sono le differenti modalità assunte dal carattere, mentre n_j e f_j sono rispettivamente la frequenza assoluta e relativa della j-esima modalità x_j. Se si conosce la distribuzione di frequenze di un carattere X suddiviso in K classi del tipo $[x_{j-1}, x_j[$ con $j = 1, \ldots, K$ possiamo approssimare la media aritmetica del carattere con la seguente espressione:

$$M_a = \sum_{j=1}^{K} c_j n_j \tag{2.3}$$

dove $c_j = \frac{x_{j-1}+x_j}{2}$ è il valore centrale della classe j-esima e n_j è la corrispondente frequenza assoluta.

Definizione 2.14. Si definisce **media aritmetica ponderata** di un carattere quantitativo X con K modalità $x_1, \ldots, x_K$ e rispettivi pesi $p_1, \ldots, p_K$, il rapporto

$$M_p = \frac{\sum_{j=1}^{K} x_j p_j}{\sum_{j=1}^{K} p_j}. \tag{2.4}$$

Proposizione 2.15. *La media aritmetica possiede un certo numero di proprietà matematiche, tutte dimostrabili in maniera elementare.*

1. *La somma delle differenze tra i valori delle x_i e la loro media aritmetica è pari a zero. In tal senso la media aritmetica può essere considerata in una posizione centrale rispetto ai valori del carattere X:*

$$\sum_{i=1}^{n} (x_i - M_a) = 0. \tag{2.5}$$

2. *La media aritmetica minimizza la somma degli scarti quadratici dei valori x_i da una costante $c \in \mathbb{R}$, cioè la funzione*

$$f(c) = \sum_{i=1}^{n} (x_i - c)^2 \qquad \text{è minima per} \qquad c = M_a. \tag{2.6}$$

3. *Siano dati $a, b \in \mathbb{R}$ e la distribuzione di un carattere X con media $M_a(X)$. Posto $Y = aX + b$, si ha $M_a(Y) = aM_a(X) + b$.*

2.4 Indici di variabilità

Esiste anche un'altra tipologia molto importante di indici statistici, atta ad analizzare la **variabilità** di una distribuzione, ossia la tendenza delle unità di un collettivo ad assumere modalità diverse del carattere. È importante

che un indice di questo tipo assuma il suo valore minimo se e solo se tutte le unità della distribuzione presentano uguale modalità del carattere; il suo valore dovrà poi aumentare al crescere della diversità tra le modalità assunte.

Definizione 2.16. La **varianza** σ^2 di un insieme di valori $x_1, \ldots, x_n$ di una variabile X con media M_a è definita come la media degli scarti quadratici dei valori x_i dalla media aritmetica:

$$\sigma^2 = \sigma^2(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_a)^2. \tag{2.7}$$

Se conosciamo la distribuzione di frequenze di una variabile X con K modalità distinte $x_1, \ldots, x_K$, per la varianza vale la seguente rappresentazione:

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{K}(x_j - M_a)^2 n_j = \sum_{j=1}^{K}(x_j - M_a)^2 f_j \tag{2.8}$$

dove M_a è la media aritmetica, mentre n_j e f_j sono rispettivamente le frequenze assolute e relative della j-esima modalità x_j.

Proposizione 2.17. *La varianza verifica due importanti proprietà.*

1. *Si ha*
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - M_a^2. \tag{2.9}$$

2. *Siano dati $a, b \in \mathbb{R}$ e un carattere X con media $M_a(X)$ e varianza $\sigma^2(X)$. Posto $Y = aX + b$, si ha $\sigma^2(Y) = a^2\sigma^2(X)$.*

Definizione 2.18. Si definisce **deviazione standard** o **scarto quadratico medio** di un carattere X la radice quadrata della sua varianza: $\sigma = \sqrt{\sigma^2}$.

Osservazione 2. Come indice di variabilità, la varianza ha il difetto di non possedere la stessa unità di misura dei valori della distribuzione (ne possiede infatti il quadrato). Per tale motivo è preferibile utilizzare come indice di variabilità la deviazione standard appena definita.

Esempio 2.19. Si consideri la Tabella 2.5. Essa riporta due distribuzioni percentuali relative a un medesimo carattere X osservato su due collettivi diversi. Indicate con $M_a(1)$ e $M_a(2)$ le rispettive medie aritmetiche, si ha: $M_a(1) = M_a(2) = 0$. Si osservi che nella seconda distribuzione le unità appaiono più concentrate intorno al valore medio, assumendo un massimo in corrispondenza di esso, mentre per la prima distribuzione ciò non accade. Anzi, la distribuzione presenta due massimi in corrispondenza dei valori -2 e 2, mentre il valore 0 è addirittura uno dei valori meno frequenti.

È qui che entra in gioco la varianza: utilizzando la formula (2.8) nel caso

X	Distrib. 1	Distrib. 2
-4	$1,2$	$1,2$
-3	$12,2$	$6,1$
-2	$24,4$	$12,2$
-1	$11,0$	$18,3$
0	$2,4$	$24,4$
1	$11,0$	$18,3$
2	$24,4$	$12,2$
3	$12,2$	$6,1$
4	$1,2$	$1,2$

Tabella 2.5: Esempio di distribuzioni con stessa media aritmetica

delle frequenze relative si ottiene $\sigma^2(1) = 4,75$ e $\sigma^2(2) = 2,82$. Con questi indici è dunque possibile affermare con certezza che la variabilità della prima distribuzione è maggiore rispetto a quella della seconda.

2.5 Frequenze congiunte e tabelle a doppia entrata

Procediamo ora all'analisi congiunta di due caratteri X e Y osservati su uno stesso collettivo di n unità. Se da un lato possiamo utilizzare la

distribuzione unitaria doppia, come definita nella Definizione 2.10, dall'altro risulta essere molto utile un tipo di rappresentazione più sintetica come la distribuzione di frequenze congiunte attraverso la tabella a doppia entrata.

Definizione 2.20. Dati due caratteri X e Y, rispettivamente con H e K modalità $x_1, \ldots, x_H$ e $y_1, \ldots, y_K$, definiamo **tabella a doppia entrata** la Tabella 2.6 il cui corpo centrale è costituito da una matrice di ordine HxK il cui generico elemento n_{ij}, per $i = 1, \ldots, H$ e $j = 1, \ldots, K$, rappresenta la **frequenza congiunta**, ossia la frequenza assoluta delle unità che presentano congiuntamente la modalità x_i di X e la modalità y_j di Y.
La colonna e la riga dei totali sono dette **distribuzioni marginali** e corrispondono esattamente alle distribuzioni di frequenze semplici dei due caratteri. In particolare, la colonna del totale è la distribuzione semplice del carattere X e il generico termine $n_{i.}$ indica la frequenza assoluta delle unità che presentano la modalità x_i; analogamente la riga del totale indica la distribuzione semplice del carattere Y e il generico termine $n_{.j}$ indica la frequenza assoluta delle unità che presentano la modalità y_j.

		Y					
		y_1	...	y_j	...	y_K	**Totale**
	x_1	n_{11}	...	n_{1j}	...	n_{1K}	$n_{1.}$
	...	...	...	...	...	...	...
X	x_i	n_{i1}	...	n_{ij}	...	n_{iK}	$n_{i.}$
	...	...	...	...	...	...	...
	x_H	n_{H1}	...	n_{Hj}	...	n_{HK}	$n_{H.}$
	Totale	$n_{.1}$	...	$n_{.j}$	...	$n_{.K}$	n

Tabella 2.6: Tabella a doppia entrata

Definizione 2.21. Le righe e le colonne interne alla tabella a doppia entrata identificano le cosiddette **distribuzioni condizionate**. In particolare,

facendo riferimento alla Tabella 2.6, si consideri la distribuzione data dalla generica riga i-esima: $n_{i1}, \ldots, n_{ij}, \ldots, n_{iK}$. Essa è detta distribuzione condizionata della Y rispetto alla modalità x_i di X.

Osservazione 3. In una tabella a doppia entrata riferita a due caratteri rispettivamente con H e K modalità, osservati su un collettivo di n unità, è immediato verificare le seguenti relazioni:

$$n_{i.} = \sum_{j=1}^{K} n_{ij} \quad \forall i = 1, \ldots, H; \qquad n_{.j} = \sum_{i=1}^{H} n_{ij} \quad \forall j = 1, \ldots, K; \tag{2.10}$$

$$n = \sum_{i=1}^{H} \sum_{j=1}^{K} n_{ij} = \sum_{i=1}^{H} n_{i.} = \sum_{j=1}^{K} n_{.j}.$$

Definizione 2.22. Anche nel caso delle tabelle a doppia entrata si può parlare di **frequenze relative** o **percentuali**, dove il generico elemento interno alla tabella a doppia entrata è espresso, rispettivamente, da $f_{ij} = n_{ij}/n$ e da $p_{ij} = f_{ij} \cdot 100$. La **distribuzione marginale relativa di** X si ottiene come $n_{i.}/n$ per $i = 1, \ldots, H$ e fornisce la distribuzione di frequenze relative semplici del carattere X. Lo stesso vale per il carattere Y, per il quale la distribuzione marginale relativa si ottiene come $n_{.j}/n$ per $j = 1, \ldots, K$.

Definizione 2.23. La **distribuzione relativa condizionata di** X **rispetto alla modalità** y_j **di** Y si ottiene come $n_{ij}/n_{.j}\, \forall i = 1, \ldots, H$.
Viceversa la **distribuzione relativa condizionata di** Y **rispetto alla modalità** x_i **di** X si ottiene come $n_{ij}/n_{i.}\, \forall j = 1, \ldots, K$.
Queste distribuzioni vengono anche dette rispettivamente **profili colonna** e **profili riga** della tabella a doppia entrata e possono essere rappresentate come in Tabella 2.7.

Esempio 2.24. Viene considerato un collettivo di $n = 21$ persone. Su di esso si analizzano due caratteri X e Y, rispettivamente X=*Sesso* nelle modalità x_1=Maschio e x_2=Femmina e Y=*Colore degli occhi* nelle modalità y_1=Marrone, y_2=Verde, y_3=Azzurro e y_4=Nero. I risultati ottenuti da questa indagine sono riportati nella Tabella 2.8, tabella a doppia entrata delle

Distribuzione relativa condizionata della X rispetto alla modalità y_j						Distribuzione relativa condizionata della Y rispetto alla modalità x_i					
x_1	...	x_i	...	x_H	Tot.	y_1	...	y_j	...	y_K	Tot.
$n_{1j}/n_{.j}$	...	$n_{ij}/n_{.j}$	...	$n_{Hj}/n_{.j}$	1	$n_{i1}/n_{i.}$	...	$n_{ij}/n_{i.}$	...	$n_{iK}/n_{i.}$	1

Tabella 2.7: Profili colonna e profili riga

frequenze congiunte n_{ij} per $i = 1, 2$ e $j = 1, 2, 3, 4$. Osservando tale tabella, si nota per esempio che ci sono 5 maschi con gli occhi marroni, 4 femmine con gli occhi azzurri, ecc.

In riferimento alla Definizione 2.21, si osserva che la prima riga della Tabel-

		Y				
		y_1	y_2	y_3	y_4	**Totale**
X	x_1	5	2	3	1	11
	x_2	3	1	4	2	10
	Totale	8	3	7	3	21

Tabella 2.8: Esempio di tabella a doppia entrata

la 2.8 rappresenta la distribuzione condizionata del carattere Y rispetto alla modalità x_1=Maschio di X, così si esamina esclusivamente il carattere *Colore degli occhi* fra gli individui di sesso maschile. Analogamente se si seleziona una colonna viene fissato un particolare colore degli occhi e si esamina fra gli individui che hanno gli occhi di quel colore quanti sono i maschi e quante le femmine.

Ricordando la Definizione 2.23, dividendo ogni riga per il totale di riga si ottengono i profili riga riportati in Tabella 2.9. Così ad esempio $0,1$ è la frazione dei 10 individui di sesso femminile che hanno gli occhi verdi. Analogamente per i profili colonna si ottiene la Tabella 2.10. Dunque $0,625$ è la frazione degli 8 individui con gli occhi marroni che sono maschi.

		Y				
		y_1	y_2	y_3	y_4	**Totale**
X	x_1	0,455	0,182	0,273	0,090	1
	x_2	0,3	0,1	0,4	0,2	1

Tabella 2.9: Esempio di profili riga

		Y			
		y_1	y_2	y_3	y_4
X	x_1	0,625	0,667	0,429	0,333
	x_2	0,375	0,333	0,571	0,667
	Totale	1	1	1	1

Tabella 2.10: Esempio di profili colonna

2.6 Associazione fra due caratteri

La ricerca scientifica non si limita alla descrizione dei singoli fenomeni, ognuno considerato indipendentemente dagli altri. Essa si interessa anche, e soprattutto, dell'analisi delle relazioni che ognuno di questi può avere con gli altri.

Definizione 2.25. Si parla di **dipendenza logica** tra due o più caratteri quando tra questi sono note a priori relazioni di causa ed effetto. Si parla invece di **indipendenza logica** quando si suppone a priori che tra questi non possa sussistere alcuna relazione di causa ed effetto.
A posteriori si potrà poi parlare di **dipendenza** o **indipendenza statistica** a seconda che la conoscenza della modalità di uno dei due caratteri possa migliorare oppure no la previsione della modalità dell'altro.

Esempio 2.26. Per esempio, a priori sappiamo che la statura di un uomo dipende dall'età, dall'alimentazione, dal patrimonio genetico dei genitori, ma certamente non dal genere musicale preferito.

Nella ricerca dell'associazione tra caratteri si possono utilizzare due approcci: l'**analisi della dipendenza**, dove si studia come le modalità di un carattere dipendano da quelle di un altro carattere secondo un legame unidirezionale, o l'**analisi dell'interdipendenza** in cui si assume che i caratteri abbiano tutti lo stesso ruolo e che il legame tra essi sia bidirezionale.
Vedremo come questi due tipi di legame, quando realizzati in modo "perfetto" corrispondano alla nozione matematica di applicazione e di applicazione biunivoca, rispettivamente.
La tabella a doppia entrata è lo strumento più idoneo per indagare su relazioni esistenti tra le modalità di due caratteri qualitativi o quantitativi suddivisi in classi.

Definizione 2.27. Il carattere X si dirà **statisticamente indipendente** da Y se, qualunque sia la modalità con cui si manifesta il carattere Y, la distribuzione relativa condizionata di X rimane sempre la stessa, cioè i profili colonna della tabella a doppia entrata sono tutti uguali fra loro: $\forall i = 1, \ldots, H$, $n_{ij}/n_{\cdot j}$ è indipendente da j, per $j = 1, \ldots, K$.

Si può facilmente dimostrare il seguente risultato.

Proposizione 2.28. *Se X è indipendente da Y allora Y è indipendente da X. Inoltre due caratteri X e Y sono **indipendenti** se e solo se $n_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$, $\forall i = 1, \ldots, H$ e $\forall j = 1, \ldots, K$. In questo caso i profili colonna, tutti uguali fra loro, coincidono con la distribuzione relativa marginale di X; analogamente i profili riga coincidono con la distribuzione relativa marginale di Y.*

Esempio 2.29. La Tabella 2.11 si riferisce a due caratteri osservati su un collettivo di 112 unità. Per verificare l'indipendenza tra i due caratteri consideriamo le distribuzioni condizionate della X rispetto alla Y, riportate nella Tabella 2.12. Per la Definizione 2.27 risulta chiaramente che i due caratteri sono indipendenti.
Volendo verificare la Proposizione 2.28 andiamo a calcolare le cosiddette **fre-**

		Y				
		y_1	y_2	y_3	y_4	**Totale**
X	x_1	3	2	4	5	14
	x_2	6	4	8	10	28
	x_3	15	10	20	25	70
	Totale	24	16	32	40	112

Tabella 2.11: Esempio di caratteri indipendenti

		Y				
		y_1	y_2	y_3	y_4	**Totale**
X	x_1	0,125	0,125	0,125	0,125	0,125
	x_2	0,250	0,250	0,250	0,250	0,250
	x_3	0,625	0,625	0,625	0,625	0,625
	Totale	1,00	1,00	1,00	1,00	1,00

Tabella 2.12: Esempio di distribuzioni condizionate della X rispetto alla Y

quenze teoriche di indipendenza, definite da $n'_{ij} = \dfrac{n_{i.} \cdot n_{.j}}{n}$ che devono coincidere con quelle empiriche n_{ij}. Si ha infatti:

$$n'_{11} = \frac{14 \cdot 24}{112} = 3 = n_{11}, \quad n'_{21} = \frac{28 \cdot 24}{112} = 6 = n_{21}, \quad n'_{23} = \frac{28 \cdot 32}{112} = 8 = n_{23},$$

e così via. L'uguaglianza può essere verificata per tutti gli n'_{ij}.
Si osservi che in generale le frequenze teoriche sono dei valori razionali in quanto sono costruite come rapporti di valori interi. Ciò implica che solo nel caso in cui tutte le frequenze teoriche risultano essere intere si può verificare che esse sono esattamente uguali a quelle empiriche e che quindi c'è indipendenza statistica tra i due caratteri. Per questo motivo la situazione di perfetta indipendenza deve essere considerata come una situazione ideale dalla quale i caratteri possono trovarsi più o meno distanti.

Un altro tipo di relazione fra due caratteri è l'**associazione spuria**, ossia

un legame statistico che si verifica tra due caratteri logicamente indipendenti, cosicché le relazioni osservate sono solo apparenti. Si noti che talvolta, l'associazione spuria tra due caratteri può essere dovuta alla presenza di uno o più caratteri non considerati che influenzano entrambi i caratteri osservati. Analizziamo ora la situazione antitetica rispetto alla indipendenza perfetta.

Definizione 2.30. Si dice che un carattere Y **dipende perfettamente** da X quando a ogni modalità di X è associata una ed una sola modalità di Y, cioè quando in una tabella a doppia entrata $\forall i\ \exists! j : n_{ij} \neq 0$.

Osservazione 4. Se Y dipende perfettamente da X, indicati con $A = \{x_1, \ldots, x_H\}$ e $B = \{y_1, \ldots, y_K\}$ gli insiemi delle modalità assunte da X e da Y sul collettivo, rispettivamente, la legge che ad $x_i \in A$ fa corrispondere $y_j \in B$ tale che $n_{ij} \neq 0$ definisce una applicazione f di A in B. Indicando con x un generico elemento di A e con $y = f(x)$ il suo corrispondente possiamo dire che y è funzione di x nell'usuale significato matematico dell'espressione.

Definizione 2.31. Se X dipende perfettamente da Y e viceversa Y dipende perfettamente da X si parla di **interdipendenza perfetta** fra i due caratteri. In termini matematici, nella notazione introdotta nell'Osservazione 4 la funzione $y = f(x)$ di A in B è biunivoca.

Ogni situazione intermedia tra l'indipendenza e l'associazione perfetta esprime un certo grado di dipendenza o interdipendenza tra i caratteri, che sarà tanto maggiore quanto più la tabella osservata si discosta da quella di indipendenza a favore di una situazione di perfetta associazione.

2.7 Associazione fra caratteri quantitativi

Esaminiamo ora in dettaglio l'associazione fra due caratteri quantitativi. Introduciamo alcune nozioni preliminari.

Definizione 2.32. Data la distribuzione unitaria doppia di due caratteri X e Y quantitativi, si definisce **baricentro** la coppia $(M_a(X), M_a(Y))$ data dalle medie aritmetiche dei due caratteri.

Definizione 2.33. A partire dalla distribuzione unitaria doppia di due caratteri quantitativi X e Y osservati su un collettivo di n unità, si definisce **grafico di dispersione** l'insieme $\{(x_i, y_i)|i = 1, \ldots, n\}$ delle coppie di modalità dei due caratteri osservate per ogni unità del collettivo. Tali coppie possono essere rappresentate come punti di un piano cartesiano i cui assi corrispondono ai due caratteri e in cui l'origine viene poi traslata in corrispondenza del baricentro.

A partire dal grafico di dispersione si parla di **concordanza** se la maggior parte degli scostamenti dalla media è concorde, cioè se la maggior parte dei punti del grafico di dispersione appartiene al I e III quadrante. Al contrario, si parla di **discordanza** se la maggior parte degli scostamenti è discorde.

Esempio 2.34. Consideriamo le due variabili X e Y rispettivamente *Percentuale di popolazione di età maggiore o uguale a 65 anni* e *Numero di posti letto per 1000 abitanti* rilevate sulle venti regioni italiane. La Tabella 2.13

Regione	**X**	**Y**	**Regione**	**X**	**Y**
Piemonte	17,2	4,9	Marche	17,7	6,2
Valle d'Aosta	15,2	5,4	Lazio	13,7	4,6
Lombardia	14,4	5,6	Abruzzo	15,8	6,5
Trentino A. A.	14,3	6,7	Molise	16,5	4,7
Veneto	14,7	7,2	Campania	10,8	3,6
Friuli V. G.	18,6	8,1	Puglia	11,8	5,8
Liguria	20,9	7,0	Basilicata	13,9	4,1
Emilia R.	18,9	6,2	Calabria	12,8	4,6
Toscana	18,7	5,9	Sicilia	12,9	4,6
Umbria	18,2	6,0	Sardegna	12,0	4,8

Tabella 2.13: Esempio di distribuzione unitaria doppia

riporta la distribuzione unitaria doppia dei due caratteri, mentre nella Figura 2.1 l'insieme dei dati è rappresentato attraverso il grafico di dispersione. Si

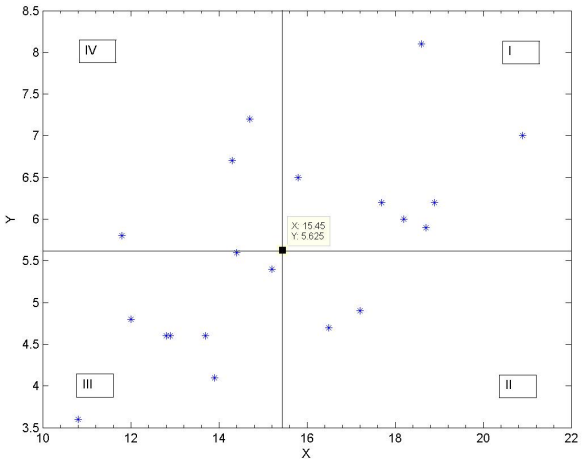

Figura 2.1: Grafico di dispersione

noti che il baricentro vale $(15.45; 5.625)$; sarà possibile fissare in questo punto la nuova origine degli assi per poter analizzare concordanza e discordanza dei due caratteri. In particolare in questo caso si verifica che prevalgono gli scostamenti concordi su quelli discordi.

Il grafico di dispersione è di per sé molto utile per evidenziare se prevale la concordanza o la discordanza fra due caratteri. È possibile tuttavia introdurre un indice che permette di effettuare la stessa analisi con un semplice calcolo, senza dover ricorrere al grafico di dispersione. Si tratta della covarianza.

Definizione 2.35. A partire dalla distribuzione unitaria doppia, la **covarianza** tra due caratteri quantitativi è definita come la media dei prodotti degli scostamenti delle variabili X e Y dalle rispettive medie:

$$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - M_a(X))(y_i - M_a(Y)). \tag{2.11}$$

Il prodotto $n\sigma_{XY}$ è detto **codevianza**.

Si può dimostrare facilmente la seguente

Proposizione 2.36. *La covarianza verifica le seguenti proprietà.*

1. $\sigma_{XY} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - M_a(X))(y_i - M_a(Y)) = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - M_a(X)M_a(Y) =$

$$= M_a(XY) - M_a(X)M_a(Y). \tag{2.12}$$

2. *Se due caratteri sono statisticamente indipendenti allora la loro covarianza è nulla.*

3. *Indicando con σ_X e σ_Y le deviazioni standard di X e Y si ha*

$$|\sigma_{XY}| \leq \sigma_X \sigma_Y. \tag{2.13}$$

Inoltre $|\sigma_{XY}| = \sigma_X\sigma_Y$ se e solo se esiste un legame perfetto di tipo lineare tra X e Y, ossia $\exists\, a, b \in \mathbb{R}$ tali che $Y = a + bX$.

Dimostrazione. La dimostrazione di *1* e *2* è piuttosto semplice e si basa sulle proprietà della media aritmetica. Per ragioni di brevità ci limitiamo a dimostrare la proprietà *3*, utilizzando la disuguaglianza di Cauchy-Schwarz in $\mathbb{R}^n$. Essa dice che se $\alpha = (\alpha_1, \ldots, \alpha_n)$ e $\beta = (\beta_1, \ldots, \beta_n)$ sono due vettori in $\mathbb{R}^n$, se con $\alpha\beta = \sum_{i=1}^{n} \alpha_i\beta_i$ indichiamo il loro prodotto scalare e con $||\alpha|| = \Big(\sum_{i=1}^{n} \alpha_i^2\Big)^{1/2}$ e $||\beta|| = \Big(\sum_{i=1}^{n} \beta_i^2\Big)^{1/2}$ le loro rispettive norme, vale la seguente disuguaglianza

$$|\alpha\beta| \leq ||\alpha|| \cdot ||\beta||. \tag{2.14}$$

Inoltre l'uguaglianza

$$|\alpha\beta| = ||\alpha|| \cdot ||\beta|| \tag{2.15}$$

vale se e solo se α e β sono linearmente dipendenti, cioè $\exists\, t \in \mathbb{R}$ tale che $\beta = t\alpha$, ovvero $\beta_i = t\alpha_i\ \forall i = 1, \ldots, n$.
Ponendo ora $\alpha_i = \dfrac{1}{\sqrt{n}}(x_i - M_a(X))$ e $\beta_i = \dfrac{1}{\sqrt{n}}(y_i - M_a(Y))\ \forall i = 1, \ldots, n$ dalla (2.14) si ottiene

$$\sigma_{XY}^2 = \frac{1}{n^2}\Big(\sum_{i=1}^{n}(x_i - M_a(X))(y_i - M_a(Y))\Big)^2 \leq$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}(x_i - M_a(X))^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - M_a(Y))^2 = \sigma_X^2\sigma_Y^2.$$

Da cui risulta: $|\sigma_{XY}| \leq \sigma_X\sigma_Y$.
Inoltre l'uguaglianza vale se e solo se $\exists\, t \in \mathbb{R}$ tale che

$$\beta_i = \frac{y_i}{\sqrt{n}} - \frac{1}{\sqrt{n}}M_a(Y) = t\frac{x_i}{\sqrt{n}} - \frac{t}{\sqrt{n}}M_a(X)$$

ovvero se e solo se $y_i = tx_i - tM_a(X) + M_a(Y)$, da cui $Y = a + bX$ con $a = M_a(Y) - tM_a(X)$ e $b = t$. □

Per la proprietà *2* della Proposizione 2.36 non vale il viceversa. Si possono infatti fornire semplici controesempi.

Esempio 2.37. Si consideri la Tabella 2.14. I caratteri X e Y sono legati da una relazione di dipendenza perfetta: $Y = \frac{1}{4}X^2$. Tuttavia la covarianza risulta essere nulla, infatti: $M_a(X) = 0$, $M_a(Y) = 5$, $M_a(XY) = 0$, quindi $\sigma_{XY} = 0$.

X	−6	−2	2	6
Y	9	1	1	9

Tabella 2.14: Esempio di dipendenza perfetta con covarianza nulla

Un difetto della covarianza è quello di dipendere dall'unità di misura delle osservazioni cosicché non è corretto confrontare il valore della covarianza calcolato su diverse distribuzioni doppie. Per ovviare tale problema è necessario introdurre un indice relativo.

Definizione 2.38. Si definisce **coefficiente di correlazione lineare** di Bravais e Pearson l'indice, atto a misurare l'interdipendenza lineare tra due caratteri, dato da:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{2.16}$$

Dalla Proposizione 2.36 segue immediatamente il seguente

Corollario 2.39. *Il coefficiente di correlazione lineare gode delle seguenti proprietà:*

1. $-1 \leq \rho_{XY} \leq 1$.
2. $\rho_{XY} = 1$ *oppure* $\rho_{XY} = -1$ *se e solo se tra i due caratteri sussiste un perfetto legame lineare del tipo* $Y = a + bX$. *In particolare, per* $b < 0$ *si ha* $\rho_{XY} = -1$ *e per* $b > 0$ *si ha* $\rho_{XY} = 1$.
3. $\rho_{XY} = 0$ *se e solo se* $\sigma_{XY} = 0$, *il che può essere dovuto all'indipendenza dei due caratteri oppure a qualche relazione di tipo non lineare (in quest'ultimo caso i due caratteri vengono detti* ***incorrelati****).*
4. *Dati* $a, b, c, d \in \mathbb{R}$ *e due caratteri quantitativi* X *e* Y, *si costruiscono le trasformazioni lineari* $X' = a + bX$ *e* $Y' = c + dY$. *Per il coefficiente di correlazione tra* X' *e* Y' *si ha*
   $\rho_{X'Y'} = \rho_{XY}$ *se i coefficienti* b *e* d *hanno stesso segno;*
   $\rho_{X'Y'} = -\rho_{XY}$ *in caso contrario.*
   In particolare, se come trasformazione lineare consideriamo proprio la standardizzazione dei caratteri: $X' = \frac{X - M_a(X)}{\sigma_X}$ *e* $Y' = \frac{Y - M_a(Y)}{\sigma_Y}$ *si ha quindi che* $\rho_{X'Y'} = \rho_{XY}$.

2.8 Regressione lineare semplice

Nell'analisi dei caratteri quantitativi si può cercare di individuare una funzione che descriva in modo sintetico le caratteristiche del legame che li unisce.

Definizione 2.40. Si definisce **modello di regressione lineare semplice** l'espressione:

$$Y = f(X; \theta) + \epsilon \tag{2.17}$$

in cui Y è la variabile dipendente, f è una generica funzione della variabile indipendente X, θ indica l'insieme dei parametri utilizzati ed ϵ rappresenta l'insieme degli effetti che altre variabili, non considerate nell'analisi, hanno

sulla Y.
Si possono utilizzare innumerevoli tipi di funzioni; una prima distinzione può essere fatta suddividendo le funzioni in lineari e non lineari nei parametri. Una **funzione lineare nei parametri** può essere scritta nel seguente modo:

$$f(X;\theta) = \theta_1 f_1(X) + \theta_2 f_2(X) + \ldots + \theta_h f_h(X) \tag{2.18}$$

di cui il polinomio di grado h del tipo: $f(X) = \theta_0 + \theta_1 X + \theta_2 X^2 + \ldots + \theta_h X^h$ è un caso particolare.
Si parla di **relazione esatta** tra le due variabili X e Y se la relazione si può esprimere come $Y = f(X;\theta)$, senza includere il termine ϵ. Al contrario, in una **relazione statistica** il valore della variabile dipendente Y non è mai univocamente determinato dal valore assunto dalla variabile indipendente X.

Osservazione 5. Il problema che si pone è quello dell'individuazione della $f(X;\theta)$ più adatta a descrivere la relazione tra le due variabili, ossia della funzione che rappresenti al meglio la nuvola dei punti osservati, data dal grafico di dispersione. Tale funzione dovrà fornire i valori teorici y_i^* più "vicini" ai valori osservati y_i della variabile Y. Questa "vicinanza" viene definita in termini geometrici, per questo a volte si preferisce parlare di **interpolazione** piuttosto che di regressione.

Nel seguito la funzione che verrà assunta come riferimento sarà un polinomio di primo grado: $f(X) = a + bX$, espressione che individua una famiglia di rette. Inoltre indicheremo con (x_i, y_i) i valori dei caratteri X e Y osservati sulla generica i-esima unità del collettivo, detti anche **valori empirici**, e con (x_i, y_i^*) i corrispondenti punti della retta. I valori $y_i^* = a + bx_i$ vengono detti **valori teorici** del carattere Y.

2.9 Metodo di interpolazione dei minimi quadrati

Il metodo di interpolazione dei minimi quadrati, che permette di individuare la retta di regressione, consiste nel minimizzare la somma dei quadrati

delle differenze tra il valore osservato y_i e il valore teorico y_i^*:

$$S(a,b) = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2. \tag{2.19}$$

Ricordando che la distanza euclidea tra due generici punti $p_1 = (x_1, y_1)$ e $p_2 = (x_2, y_2)$ è definita $d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, la distanza euclidea tra un valore osservato $p_i = (x_i, y_i)$ e il corrispondente punto sulla retta $p_i^* = (x_i, y_i^*) = (x_i, a + bx_i)$ è data da $d(p_i, p_i^*) = \sqrt{(x_i - x_i)^2 + (y_i - a - bx_i)^2} = \sqrt{(y_i - a - bx_i)^2}$.
Minimizzando la funzione $S(a,b) = \sum_{i=1}^{n} (d(p_i, p_i^*))^2$ si individueranno i due parametri a^* e b^* della retta di regressione, la quale minimizza la somma dei quadrati delle distranze euclidee tra i punti (x_i, y_i) della nuvola dei punti osservati, data dal grafico di dispersione, e i corrispondenti punti della retta (x_i, y_i^*). Più precisamente si dà la seguente

Definizione 2.41. Si chiama **retta di regressione** la retta di equazione $Y^* = a^* + b^* X$, dove la coppia (a^*, b^*) minimizza in $\mathbb{R}^2$ la funzione di due variabili $S(a,b)$ definita da (2.19).

Il problema della determinazione della retta di regressione è risolto dalla seguente

Proposizione 2.42. *I valori che minimizzano $S(a,b)$ sono dati da*

$$b^* = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}; \qquad a^* = \overline{y} - b^* \overline{x} \tag{2.20}$$

dove $\overline{x} = M_a(X)$, $\overline{y} = M_a(Y)$ e b^* è detto ***coefficiente di regressione****.*
La retta di regressione che meglio si adatta ai dati osservati è dunque data da $f(X) = a^* + b^* X$. Sostituendo le espressioni trovate si ottiene

$$f(X) = \overline{y} + \frac{\sigma_{XY}}{\sigma_X^2}[X - \overline{x}]. \tag{2.21}$$

Dimostrazione. Per determinare il minimo della funzione $S(a,b)$, si deve innanzitutto imporre la condizione $\nabla S(a,b)=0$ per la ricerca dei punti critici. Dalla condizione $\nabla S(a,b)=0$ si ottiene un sistema di due equazioni in due incognite, detto **sistema normale**:

$$\begin{cases} \dfrac{\partial S(a,b)}{\partial a} = -2\displaystyle\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \dfrac{\partial S(a,b)}{\partial b} = -2\displaystyle\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0 \end{cases} \tag{2.22}$$

Ricordiamo la Definizione 2.13 di media aritmetica nella formulazione (2.1) e utilizziamo le notazioni $\overline{x} = M_a(X)$ e $\overline{y} = M_a(Y)$. Il sistema normale (2.22) può essere così risolto

$$\begin{cases} n\overline{y} - na - bn\overline{x} = 0 \quad \text{da cui} \quad a = \overline{y} - b\overline{x} \\ \displaystyle\sum_{i=1}^{n} x_i y_i - an\overline{x} - b\sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

Nella seconda equazione sostituiamo $a = \overline{y} - b\overline{x}$ e ricordiamo la Definizione 2.16 di varianza e 2.35 di covarianza, in particolare nelle loro scritture equivalenti (2.9) e (2.12). Si avrà: $n\sigma_{XY} - bn\sigma_X^2 = 0$, da cui si può ricavare b. In definitiva si ottiene

$$\begin{cases} a^* = \overline{y} - \dfrac{\sigma_{XY}}{\sigma_X^2}\overline{x} \\ b^* = \dfrac{\sigma_{XY}}{\sigma_X^2} \end{cases}$$

Per verificare che la coppia (a^*, b^*) rappresenta un minimo per $S(a,b)$ introduciamo la matrice Hessiana $H(a,b)$ definita da

$$H(a,b) = \begin{pmatrix} \dfrac{\partial^2 S}{\partial a^2}(a,b) & \dfrac{\partial^2 S}{\partial ab}(a,b) \\ \dfrac{\partial^2 S}{\partial ba}(a,b) & \dfrac{\partial^2 S}{\partial b^2}(a,b) \end{pmatrix}$$

che deve risultare definita positiva in (a^*, b^*). Ricordiamo che se $A = (a_{ij})_{i,j=1,\dots,n}$ è una matrice reale nxn simmetrica, essa è definita positiva se e solo se $a_{11} > 0$ e $detA > 0$. Per la matrice $H(a,b)$ si ha

$$det(H(a^*,b^*)) = \begin{vmatrix} \dfrac{\partial^2 S}{\partial a^2}(a^*,b^*) & \dfrac{\partial^2 S}{\partial ab}(a^*,b^*) \\ \dfrac{\partial^2 S}{\partial ba}(a^*,b^*) & \dfrac{\partial^2 S}{\partial b^2}(a^*,b^*) \end{vmatrix} = \begin{vmatrix} 2n & 2n\overline{x} \\ 2n\overline{x} & 2\displaystyle\sum_{i=1}^{n} x_i^2 \end{vmatrix} = 4n\sum_{i=1}^{n} x_i^2 -$$

$$-4n^2\overline{x}^2 = 4n^2\sigma_X^2 + 4n^2\overline{x}^2 - 4n^2\overline{x}^2 = 4n^2\sigma_X^2 > 0.$$

Poiché $H_{11}(a^*, b^*) = 2n > 0$, la Proposizione 2.42 risulta dimostrata. □

Osservazione 6.

1. Il coefficiente di regressione, così come ρ_{XY}, è un indice che misura la dipendenza lineare tra due caratteri, con la differenza che b^* è un indice asimmetrico, cioè assume un valore diverso se si scambia il ruolo assunto dai due caratteri.

2. Il coefficiente di regressione b^* varia in $\mathbb{R} \cup \{\infty\}$ e ha come unità di misura il rapporto tra le unità di misura di Y e X. Esso indica, a meno di ϵ, di quanto varia Y in corrispondenza di una variazione unitaria di X.

3. $\sigma_X^2 > 0$ sempre, quindi il segno di b^* è determinato dal segno di σ_{XY}.

4. Se X e Y sono statisticamente indipendenti allora $b^* = 0$; tuttavia non vale il contrario poiché, come si è visto, σ_{XY} può annullarsi anche quando non vale l'indipendenza tra i caratteri. Nel caso $b^* = 0$ la retta di regressione è parallela all'asse delle ascisse e interseca l'asse delle ordinate nel punto $(0, \overline{y})$.

5. Tutti i punti del piano che soddisfano il sistema normale appartengono alla retta di regressione, quindi anche il baricentro $(\overline{x}, \overline{y})$ appartiene a tale retta, come si può facilmente verificare.

In maniera analoga a quanto svolto finora, avendo esaminato la relazione $Y = a + bX + \epsilon_1$, si può esaminare la relazione $X = c + dY + \epsilon_2$, per la quale si ottengono risultati analoghi. In particolare, i parametri che si possono ottenere con il metodo dei minimi quadrati sono

$$d^* = \frac{\sigma_{XY}}{\sigma_Y^2}, \qquad c^* = \overline{x} - d^*\overline{y}; \tag{2.23}$$

quindi la retta di regressione che rappresenta X in funzione di Y diventa

$$X^* = f(Y) = \overline{x} + \frac{\sigma_{XY}}{\sigma_Y^2}[Y - \overline{y}]. \tag{2.24}$$

Osservazione 7. Se si considera la tabella a doppia entrata di due caratteri X e Y suddivisi rispettivamente in H e K classi, le formule (2.20) e (2.23) per i coefficienti di regressione diventano

$$b^* = \frac{\sum_{i=1}^{H}\sum_{j=1}^{K} x_i y_j n_{ij} - n\overline{xy}}{\sum_{i=1}^{H} x_i^2 n_{i.} - n\overline{x}^2}, \qquad d^* = \frac{\sum_{i=1}^{H}\sum_{j=1}^{K} x_i y_j n_{ij} - n\overline{xy}}{\sum_{j=1}^{K} y_j^2 n_{.j} - n\overline{y}^2} \tag{2.25}$$

in cui x_i e y_j sono rispettivamente i valori centrali dell'i-esima classe di X e della j-esima classe di Y.

2.10 Varianza spiegata e varianza residua

Attraverso i coefficienti di regressione è possibile introdurre un nuovo indice che consente di spiegare ulteriormente il legame fra i due caratteri X e Y. È opportuno premettere la seguente

Proposizione 2.43. *I coefficienti di regressione verificano le seguenti proprietà:*

1. *b^* e d^* sono legati dalla relazione: $b^* = d^*\dfrac{\sigma_Y^2}{\sigma_X^2}$. Da ciò segue che se $\sigma_Y^2 = \sigma_X^2 = 1$ allora $b^* = d^*$.*

2. *b^* e d^* hanno sempre lo stesso segno, che è quello della covarianza. Così, se $\sigma_{XY} > 0$ le due rette di regressione sono entrambe crescenti, se $\sigma_{XY} < 0$ sono entrambe descescenti oppure se $\sigma_{XY} = 0$ sono orizzontali rispetto al proprio asse. In quest'ultimo caso le due rette sono tra loro perpendicolari e questo in coerenza con la Proposizione 2.28.*

3. *Il prodotto tra i due coefficienti di regressione è uguale al quadrato del coefficiente di correlazione: $b^*d^* = \rho_{XY}^2$.*

Definizione 2.44. Si definisce **indice di determinazione** il quadrato del coefficiente di correlazione ρ_{XY}^2.

Proposizione 2.45. *Per l'indice di determinazione vale la seguente rappresentazione*

$$\rho^2_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \overline{y})^2}{\sigma^2_Y}. \tag{2.26}$$

Inoltre, poiché $\overline{y} = \overline{y}^*$, la (2.26) può essere riscritta come

$$\rho^2_{XY} = \frac{\sigma^2_{Y^*}}{\sigma^2_Y}. \tag{2.27}$$

Dimostrazione. Dalla Proposizione 2.43 si ha

$$\rho^2_{XY} = b^* d^* = (b^*)^2 \frac{\sigma^2_X}{\sigma^2_Y}. \tag{2.28}$$

Poiché $Y^* = a^* + b^* X$, dalla Proposizione 2.17 si ha $\sigma^2_{Y^*} = (b^*)^2 \sigma^2_X$. Dalla (2.28) si ottiene quindi $\rho^2_{XY} = \frac{\sigma^2_{Y^*}}{\sigma^2_Y}$ che prova la (2.27).
Per ottenere la (2.26) basta far vedere che $\overline{y}^* = \overline{y}$; dalla Proposizione 2.15 e dalla (2.20) si ha infatti $M_a(Y^*) = M_a(a^* + b^* X) = a^* + b^* \overline{x} = \overline{y}$. □

Proposizione 2.46. *La varianza totale σ^2_Y può essere scritta come somma di due componenti:*

$$\sigma^2_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \overline{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2. \tag{2.29}$$

Dimostrazione. Si ha

$$\sigma^2_Y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^* + y_i^* - \overline{y})^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \overline{y})^2 + \frac{2}{n}\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \overline{y}).$$

Basta far vedere che

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)(y_i^* - \overline{y}) = 0.$$

Utilizzando le (2.9), (2.12) e (2.20) si ha infatti

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

$$= \frac{a^*}{n}\sum_{i=1}^{n} y_i + \frac{b^*}{n}\sum_{i=1}^{n} x_i y_i - M_a[(Y^*)^2] - (\overline{y})^2 + (\overline{y}^*)^2 =$$

$$= \overline{y}a^* + b^* M_a(XY) - \sigma^2_{Y^*} - (\overline{y})^2 = \overline{y}a^* + b^*\sigma_{XY} + b^*\overline{x}\,\overline{y} -$$

$$-(b^*)^2\sigma^2_X - (\overline{y})^2 = \overline{y}(a^* + b^*\overline{x} - \overline{y}) = 0.$$

□

Osservazione 8. In riferimento alla Proposizione 2.46 la **varianza totale** è la media delle distanze al quadrato tra i punti osservati y_i e i punti sulla retta parallela all'asse delle ascisse $Y = \overline{y}$. Essa risulta essere somma di due termini, la varianza spiegata e la varianza residua. La **varianza spiegata** o **varianza di regressione** $\sigma^2_{Y^*}$ è la varianza spiegata dalla retta di regressione ed è la media della distanze al quadrato tra i valori y_i^* e la retta costante $Y = \overline{y}$. Infine, la **varianza residua** è una media delle distanze al quadrato tra i punti osservati y_i e quelli della retta di regressione y_i^*.
Dunque la varianza totale riflette la variabilità dei valori della Y quando non viene utilizzata l'informazione data dai valori della X; al contrario, la varianza residua $(\sigma^2_Y - \sigma^2_{Y^*})$ esprime ciò che rimane della variabilità della Y dopo aver utilizzato le informazioni della X mediante il modello di regressione lineare semplice. Infine, la varianza spiegata $(\sigma^2_Y - var.residua)$ esprime la riduzione della variabilità totale della Y associata all'uso della X nella previsione della Y.

Possiamo riassumere queste osservazioni nella seguente

Proposizione 2.47. *L'indice di determinazione è una misura relativa della riduzione della variabilità di Y, ottenuta rapportando la varianza spiegata su quella totale (come visto nella Proposizione 2.45):*

$$\rho^2_{XY} = \frac{\sigma^2_{Y^*}}{\sigma^2_Y} = \frac{varianza \quad spiegata}{varianza \quad totale} = 1 - \frac{varianza \quad residua}{varianza \quad totale}. \tag{2.30}$$

Osservazione 9. Dalla (2.30) si ha che $\rho^2_{XY} \in [0, 1]$. Assume il valore 1 quando la relazione statistica è perfetta, cioè quando tutti i valori osservati di Y appartengono alla retta di regressione cosicché tutti gli scostamenti $|y_i - y_i^*|$

sono nulli. Da ciò segue che la varianza residua è nulla e che $\sigma_Y^2 = \sigma_{Y^*}^2$, dunque $\rho_{XY}^2 = 1$.
Il caso opposto avviene quando la relazione individuata dalla retta di regressione non è di alcun aiuto alla riduzione della variabilità totale. Quindi la variabilità residua coincide con quella totale e così $\rho_{XY}^2 = 0$. In questo caso la nuvola dei punti del grafico di dispersione non individua alcuna relazione lineare tra i due caratteri.
Generalmente nei casi reali $\rho_{XY}^2 \in]0, 1[$ e quanto più è vicino a 1, tanto maggiore è il grado di relazione lineare presentato dalle osservazioni e quindi tanto più la retta di regressione "spiega" la variabilità totale.

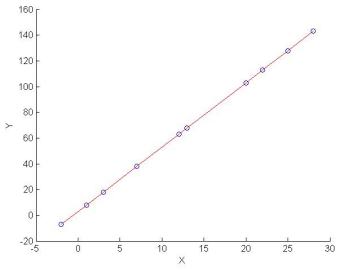

Figura 2.2: Massima correlazione positiva: $\rho_{XY}^2 = 1$

Esempio 2.48. Per capire meglio l'Osservazione 9 si osservino le Figure 2.2, 2.3, 2.4 e 2.5. Esse riportano alcuni esempi di grafici corrispondenti a diversi valori dell'indice di determinazione ρ_{XY}^2. In Figura 2.2 è stata graficata la retta di regressione nel caso di massima correlazione positiva tra i valori osservati. In questa situazione vale: $\rho_{XY} = 1 \Rightarrow \rho_{XY}^2 = 1$, indicando che la relazione statistica è perfetta; la retta di regressione passa infatti per tutti i punti osservati. In Figura 2.3 è riportato il grafico di incorrelazione tra i due caratteri: $\rho_{XY}^2 = 0$. Come si può notare al variare di X il valore di Y previsto dalla retta di regressione rimane costante; si osservi che infatti $b^* = 0$. Nel grafico di Figura 2.4 si ha una situazione intermedia: le osservazioni sono correlate positivamente ($\rho_{XY} = 0,64$), ma la retta di regressione spiega solo il

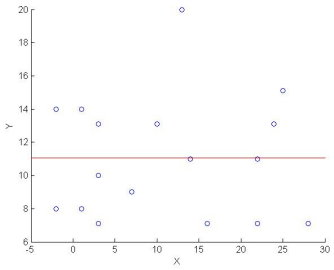

Figura 2.3: Incorrelazione: $\rho^2_{XY} = 0$

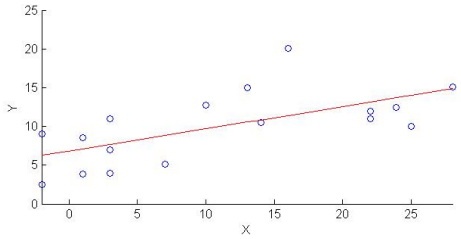

Figura 2.4: Correlazione positiva: $\rho^2_{XY} = 0,41$

41% della variabilità totale (perchè $\rho^2_{XY} = 0,41$). Infine, nel grafico di Figura 2.5 la retta di regressione ha inclinazione negativa in quanto i caratteri sono correlati negativamente ($\rho_{XY} = -0,98$); inoltre però tale retta spiega il 96% della variabilità totale (infatti $\rho^2_{XY} = 0,96$).

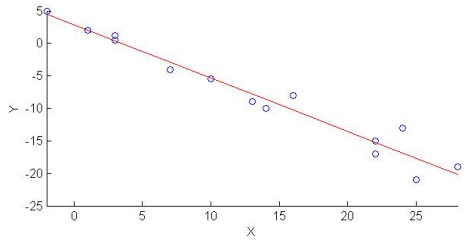

Figura 2.5: Correlazione negativa: $\rho^2_{XY} = 0,96$

Bibliografia

[1] Berengo F., *L'uso delle tecnologie nella didattica della matematica: l'esperienza dell'ITSOS "Marie Curie"*, Form@re Open journal per la formazione in rete, **38** (7), 2005, *http://formare.erickson.it/wordpress/it/2005/luso-delle-tecnologie-nella-didattica-della-matematica-lesperienza-dellitsos-marie-curie/*, 21 gennaio 2013.

[2] Borra S., Di Ciaccio A., *Statistica. Metodologie per le scienze economiche e sociali*, II edizione, Milano, McGraw-Hill, 2008.

[3] Di Ciaccio A., Borra S., *Introduzione alla statistica descrittiva*, Milano, McGraw-Hill, 1996.

[4] Di Martino P., Fiorentino G., Zan R., *Il progetto ELTP: dai test a scelta multipla ai percorsi individualizzati*, TD Tecnologie Didattiche, **19** (3), 2011, pp.163-169, *http://www.tdjournal.itd.cnr.it/files/pdfarticles/PDF54/5_TD54_DiMartino_Fiorentino_Zan.pdf*, 21 gennaio 2013.

[5] Mecatti F., *Statistica di base. Come, quando, perchè*, Milano, McGraw-Hill, 2010.

[6] Suraci D., *Ipertesti: una nuova rivoluzione culturale?*, Tracciati Rivista alla ricerca della scuola, **2** (2), 1997, *http://www.graffinrete.it/tracciati/storico/tracciati2/sur1.htm*, 21 gennaio 2013.

[7] *E-learning*, Wikipedia L'enciclopedia libera, *http://it.wikipedia.org/wiki/E-learning*, 21 gennaio 2013.

[8] Zan R., *Mortalità universitaria e mortalità matematica*, Tracciati Rivista alla ricerca della scuola, **2** (2), 1997, *http://www.graffinrete.it/tracciati/storico/tracciati2/mort.htm*, 21 gennaio 2013.

Ringraziamenti

Il ringraziamento più sentito va indubbiamente a mia madre Dilma, grande donna, fonte inesauribile di forza e mio indispensabile punto di riferimento. Ringrazio poi Daniele ed Elia per avermi sempre sostenuta in questi anni, il primo con l'efficacia della ragione e l'intensità dell'amore, il secondo con la spiazzante potenza di un sorriso.
Inoltre, pur non potendoli citare singolarmente, ringrazio tutti coloro che hanno sempre creduto in me, anche nei momenti in cui il percorso si faceva più intricato e difficile.
Infine ringrazio moltissimo la professoressa Emanuela Caliceti per avermi guidato nella stesura di questa tesi con preziosi consigli e grande disponibilità e il professor Alessandro Gimigliano per il fondamentale aiuto nella realizzazione web del mio ipertesto.