CAPITOLO 1

Elementi di statistica descrittiva

La *Statistica* è una disciplina scientifica che si occupa della raccolta, analisi ed interpretazione dei dati ottenuti da osservazioni sperimentali, di norma caratterizzate da notevole variabilità.

A linee molto generali, la statistica si divide in *metodologica*, che si occupa dei metodi per la raccolta dei dati, *descrittiva*, che si occupa della loro descrizione e sintesi e *inferenziale* che si occupa delle conclusioni che si possono trarre dall'analisi dei dati statistici.

In queste brevi note assumeremo che i dati siano già stati raccolti e ci occuperemo quasi esclusivamente di statistica descrittiva, con qualche breve accenno agli aspetti inferenziali.

Avvertenza: i dati numerici riportati negli esempi sono tratti da fonti non sempre attendibili (wikipedia, giornali, TV, memoria degli autori ecc.) e devono essere pertanto presi come base di lavoro per una discussione puramente ipotetica e non necessariamente legata alla realtà dei fatti.

1.1. Brevissima introduzione alla statistica

La raccolta di dati sulla popolazione, sui terreni e sulle proprietà ha avuto origine nella notte dei tempi, ancor prima dell'invenzione della scritture. Ad esempio si fa riferimento ad un censimento nel racconto evangelico della nascita di Gesù. Si hanno notizie storiche che nel 1589 il termine **statistica** veniva usato per indicare il complesso di conoscenze che descrivono le qualità che caratterizzano uno Stato e gli elementi che lo compongono. La nascita della statistica, nella moderna accezione del termine, viene da molti considerata in coincidenza con la pubblicazione, nel 1662 a Londra, del saggio di John Graunt sulle "osservazioni eseguite sui bollettini della mortalità, con riguardo al governo, alla religione, al commercio, allo sviluppo, al clima, alle malattie e ai vari mutamenti della città".

Oggigiorno, in Italia è costantemente ripetuto dai vari notiziari che l'ISTAT raccoglie e pubblicizza[1] dati sulla popolazione (nascite, decessi ecc.) e sull'economia (inflazione, disoccupazione, prodotto interno lordo ecc.).

La statistica è normalmente interessata ad ottenere informazioni su un insieme completo di soggetti detto **popolazione**. I soggetti di una popolazione soggetta ad indagine statistica non devono necessariamente essere persone, ma possono benissimo essere animali, piante, rocce, manufatti, libri, bagagli di conoscenze, giorni, momenti temporali, osservazioni ripetute di un fenomeno eccetera. Esempi di popolazioni ed indagini statistiche sono elencate nella seguente tabella:

[1] www.istat.it

Popolazione	**Indagine statistica**
Studenti della Sapienza	CFU acquisiti nell'anno accademico
Docenti della Sapienza	Esami verbalizzati nell'anno accademico
Pecore australiane	Efficacia del vaccino contro l'antrace
Clienti supermercato	Gusti alimentari
Distributori carburante	Prezzo del gasolio
Residenti nel Lazio	Gruppo sanguigno e fattore Rh
Abitanti di Londra nel 1636	Decessi totali e decessi per peste
Conoscenze dello studente	Test di autovalutazione
Giorni dell'anno	Temperatura massima nel centro di Roma
Lanciatori di giavellotto	Lunghezza del lancio
Squadre di calcio	Reti segnate in campionato

Nei primi due esempi della tabella, non è difficile ricavare da Infostud tutte le informazioni necessarie, le quali scritte su carta a caratteri normali formerebbero un mappazzone illeggibile ed indigesto. Al fine di fornire dei dati chiari, di immediata comprensione ed in grado di dare al lettore un'idea generale del fenomeno osservato, nel corso del tempo sono state selezionate un certo numero di tecniche e procedure, molte delle quali oramai entrate nel lessico comune: si dirà (i dati numerici sono di fantasia) ad esempio che:

(1) gli studenti della Sapienza conseguono in media 35 CFU all'anno;
(2) la mediana dei CFU conseguiti dagli studenti della Sapienza è 30;
(3) la deviazione standard del numero di esami verbalizzati dai docenti della Sapienza è 17;
(4) il Prof. Carogna supera il terzo quartile in quanto a numero di studenti bocciati agli esami.

Ritroveremo e spiegheremo i concetti di media, mediana, deviazione standard e quartile nelle prossime sezioni.

A volte la popolazione è troppo grande per poter effettuare una raccolta dati esaustiva. In tutti questi casi si cerca di imparare qualcosa scegliendo ed esaminando un sottoinsieme della popolazione, detto **campione statistico**.

Ad esempio, nel 2008, per capire quanti articoli di ricerca avevano in media i matematici delle università Italiane, l'Unione Matematica Italiana effettuò un'indagine statistica su un campione di circa il 5% della popolazione di riferimento scelto casualmente. I risultati ottenuti risultarono molto vicini a quelli calcolati successivamente sull'intera popolazione.

La scelta del campione è compito alquanto delicato ed una scelta errata può facilmente portare a risultati completamente falsati. Anche la decisione su quanto deve essere grande un campione affinché dia risultati statisticamente significativi è cosa tutt'altro che banale e richiede competenze matematiche e probabilistiche.

In generale solo campioni scelti completamente a caso sono rappresentativi, ogni criterio di selezione non casuale finisce con il produrre campioni che non sono rappresentativi. Ad esempio, se per valutare l'età media della popolazione di un piccolo comune intervistiamo tutti coloro che nel corso di una giornata entrano nell'ufficio postale, molto probabilmente dai dati raccolti risulterà un'età media più alta di quella reale.

Un esempio da manuale di errata scelta del campione si è verificato nel 1936 quando la rivista *Literary Digest* condusse un sondaggio elettorale reclutando un campione di 10 milioni di cittadini dalle liste del registro automobilistico e dell'elenco telefonico degli Stati Uniti. Solo 2,3 milioni di persone parteciparono all'indagine. Secondo il Digest, le elezioni presidenziali sarebbero state certamente vinte dal repubblicano Landon. Contemporaneamente, lo statistico George Gallup predisse invece la vittoria del democratico Roosevelt utilizzando un campione di 50.000 persone.

Le elezioni furono vinte da Roosevelt e Literary Digest chiuse i battenti dopo due anni. La lezione di Gallup fu fondamentale negli anni a venire, dimostrando la scarsa importanza della dimensione del campione statistico rispetto alla sua composizione, che dev'essere necessariamente casuale e probabilistica. Servendosi infatti degli elenchi del telefono e del registro automobilistico, il Digest aveva finito per intervistare troppi repubblicani (all'epoca mediamente più ricchi e più facilmente in possesso di un'utenza telefonica o di una automobile) e di conseguenza aveva sottostimato l'elettorato democratico.

Il sondaggio elettorale, essendo basato su campioni ridotti ed analisi probabilistiche, entra a pieno titolo nell'ambito della statistica inferenziale. La **statistica descrittiva** si occupa invece dell'analisi dei dati osservati, prescindendo da qualsiasi modello probabilistico. Lo scopo basilare della statistica descrittiva è di ridurre il volume dei dati osservati, esprimendo i dati primari dell'informazione contenuta per mezzo di grafici e indicatori numerici che li descrivono; possono essere fatte indagini comparative e si può calcolare il livello la corrispondenza dei dati sperimentali a un certo modello teorico.

Abbiamo già parlato del concetto di popolazione di una indagine statistica; ogni singolo elemento della popolazione viene talvolta detto **unità osservata**. Un altro concetto fondamentale in statistica è quello di **variabile**, intesa come una caratteristica di una unità osservata. Sono esempi di variabili il gruppo sanguigno, che assume uno dei 4 valori 0,A,B,AB, l'età ed il peso di una persona, la quotazione di un titolo azionario eccetera. Le variabili si classificano in: **numeriche**, **ordinali** e **categoriali**.

Le variabili numeriche sono quelle descritte da un numero reale; una variabile numerica si dice **discreta** se i possibili valori appartengono ad una successione crescente di numeri reali $\cdots < y_i < y_{i+1} < \cdots$, con $i \in \mathbb{Z}$. Una variabile numerica si dice **continua** se non è discreta.

Quindi una variabile che assume come possibili valori le potenze intere di 10 è discreta, mentre una variabile che assume come valori i numeri compresi tra 0 e 1 è continua. Ogni variabile numerica che assume al più un insieme finito di valori è chiaramente discreta.

Esempi di variabili numeriche discrete sono l'età (in anni) e la temperatura (in gradi Celsius); sono invece variabili numeriche continue la velocità dei venti e la magnitudo dei terremoti. In pratica ogni misurazione di una variabile numerica, essendo soggetta ad approssimazioni, risulta discretizzata.

Le variabili ordinali sono quelle che assumono valori, che pur non essendo numerici, possono essere ordinati dal più piccolo al più grande. Esempi di variabili ordinali sono i giudizi espressi sulla qualità di un lavoro (pessimo, insufficiente, sufficiente, discreto, buono, ottimo, eccellente), sul grado di soddisfazione (decisamente no, più no che sì , più sì che no, decisamente si), sul livello di dolore percepito a seguito di

una iniezione (nullo, lieve, forte, insopportabile), sull'aspetto estetico di una persona (brutta, normale, carina, bella, bellissima) eccetera.

Per consentire analisi statistiche più raffinate, talvolta le variabili ordinali vengono trasformate in variabili numeriche discrete. Ad esempio, è tradizione consolidata assegnare ai giudizi i voti pessimo=4, insufficiente=5, sufficiente=6, discreto=7, buono=8, ottimo=9, eccellente=10.

Le variabili categoriali sono quelle che non rientrano nei casi precedenti, come ad esempio il gruppo sanguigno, il sesso, il colore, la squadra del cuore, l'attore preferito, il partito votato alle ultime elezioni.

Esempi di variabili statistiche

Popolazione	**Variabile (tipo)**
Calciatori	reti segnate in Serie A (numerica discreta)
Bambini nati nell'ospedale X	Peso in kg (numerica continua)
Giorni del mese	livello di dolore del Sig. B. (ordinale)
Tifosi del Frosinone	sesso (categoriale)

Esercizi.

Esercizio 1.1. Questo esercizio, essendo basato su un episodio realmente accaduto, avvalora l'ipotesi che molti giornalisti nostrani hanno le stesse competenze statistiche di un carassio. Al telegiornale della sera viene detto che nel corso dell'ultimo anno sono morti in incidenti stradali 46 motociclisti; di questi 24 portavano il casco ed i rimanenti 22 invece erano senza casco. Il giornalista commenta i dati dicendo che portare il casco non riduce la mortalità[2]. Dire dove e perché il giornalista ha detto una cavolata.

Esercizio 1.2. Per fare un sondaggio elettorale, quale dei seguenti metodi di selezione produrrà il campione più rappresentativo:

(1) intervistare tutti i maggiorenni che escono da uno stadio al termine di una partita di calcio;
(2) intervistare tutti i maggiorenni che escono da un lussuoso ristorante in centro;
(3) usare i risultati di un sondaggio televisivo basato sulle telefonate dei telespettatori;
(4) ottenere una copia delle liste elettorali, scegliere 100 elettori a caso ed intervistarli;
(5) scegliere dei nomi dall'elenco telefonico ed intervistarli;
(6) intervistare tutti i partecipanti ad un raduno di ufologia complottistica.

Esercizio 1.3. Per capire le aspettative di vita in città, un tizio legge tutti i necrologi pubblicati nel principale quotidiano e prende nota dell'età dei deceduti. Dire se tale procedura è giusta o sbagliata.

[2]Era il TG regionale toscano di qualche tempo fa; gli autori non ricordano i numeri esatti, ma solo che i morti con il casco erano leggermente superiori a quelli senza casco.

Squadra	Pt	V	P	S	GF	GS	DR
Juventus	87	26	9	3	72	24	48
Roma	70	19	13	6	54	31	23
Lazio	69	21	6	11	71	38	33
Fiorentina	64	18	10	10	61	46	15
Napoli	63	18	9	11	70	54	16
Genoa	59	16	11	11	62	47	15
Sampdoria	56	13	17	8	48	42	6
Inter	55	14	13	11	59	48	11
Torino	54	14	12	12	48	45	3
Milan	52	13	13	12	56	50	6
Palermo	49	12	13	13	53	55	-2
Sassuolo	49	12	13	13	49	57	-8
Verona	46	11	13	14	49	65	-16
Chievo	43	10	13	15	28	41	-13
Empoli	42	8	18	12	46	52	-6
Udinese	41	10	11	17	43	56	-13
Atalanta	37	7	16	15	38	57	-19
Cagliari	34	8	10	20	48	68	-20
Cesena	24	4	12	22	36	73	-37
Parma	19	6	8	24	33	75	-42

Pt=punti,
V=vittorie,
P=pareggi,
S=sconfitte,
GF=goal fatti,
GS=goal subiti,
DR=differenza reti.

TABELLA 1. Classifica finale del campionato di calcio di Serie A 2014-15.

1.2. Rappresentazione dei dati

Le indagini statistiche si basano sulla raccolta di notevoli quantità di dati. Sarebbe però dispersivo e poco utile presentare i risultati dell'indagine pubblicando per esteso tutti i dati raccolti. Si pone quindi il problema di sintetizzare i risultati in qualche forma schematica, in modo da mettere in evidenza le informazioni principali e tralasciando quelle che, almeno per il momento, non ci interessano.

Uno dei metodi più usati per la rappresentazione dei dati, almeno quando ci interessa analizzare la loro dipendenza da pochi parametri, è mediante l'uso di tabelle. Ciascun individuo di una società evoluta viene a contatto sin dalla più tenera età con i più disparati tipi di tabelle (pagelle, tabelline Pitagoriche, orari delle lezioni ecc.) ed è del tutto superfluo aggiungere ulteriori esplicazioni: un ben noto esempio è riportato nella Tabella 1.

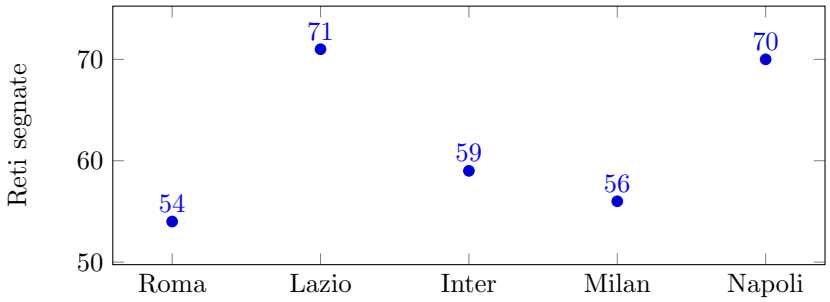
Come ulteriore esempio consideriamo un'indagine statistica sul numero di docenti in servizio nelle Università italiane nel periodo dal 2005 al 2014: possiamo presentare i risultati con la seguente tabella[3]:

[3]I dati sono presi dall'ufficio di statistica del MIUR, www.statistica.miur.it.

Anno	Prof. Ordinario	Prof. Associato	Ricercatore	Totale
2005	19274	18967	22010	60251
2006	19843	19086	23045	61974
2007	19623	18735	23571	61929
2008	18929	18256	25583	62768
2009	17880	17567	25435	60882
2010	15854	16955	24939	57748
2011	15242	16611	24596	56449
2012	14522	16143	24264	54929
2013	13890	15810	23746	53446
2014	13263	17541	21035	51839

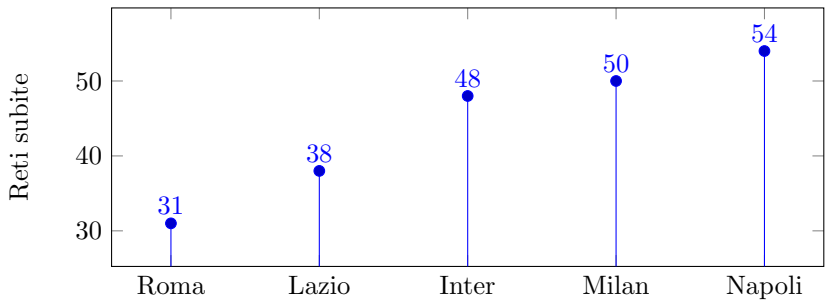
La lettura e l'interpretazione di una tabella rischia tuttavia di essere lenta e faticosa, soprattutto quando i dati numerici contenuti sono abbondanti ed una rappresentazione grafica può rendere la lettura dei dati più immediata ed efficace. Il tipo di grafica da utilizzare dipende in gran parte dal tipo di dati che si vogliono rappresentare e dal messaggio, implicito od esplicito, che si vuole trasmettere con la loro pubblicazione. I prossimi esempi descrivono i principali metodi di rappresentazione grafica.

Esempio 1.2.1. **Grafico per punti** delle reti segnate nel campionato di Serie A 2014-15.



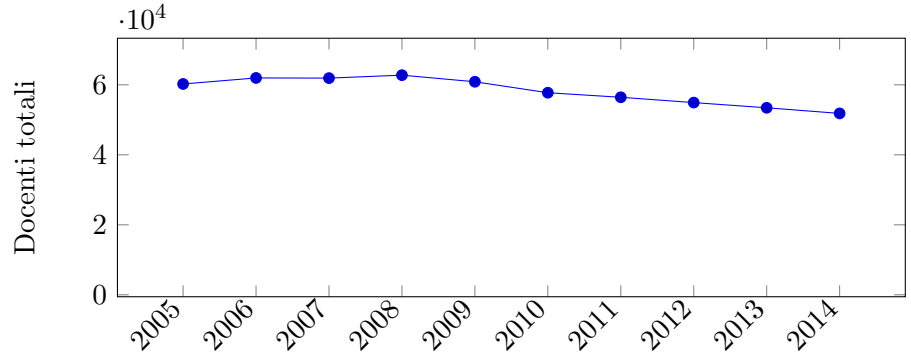
Nei grafici per punti si considera un sistema di assi cartesiani e si rappresenta le coppie di dati tra loro correlati mediante opportuni simboli grafici (cerchi, crocette, quadratini ecc.). Si noti che l'unità osservata "squadra" è di tipo non numerico; in tal caso la retta sul quale sono posizionate Roma e Lazio funge solo come supporto visivo.

Esempio 1.2.2. **Grafico a pettine** delle reti subite nel campionato di Serie A 2014-15.

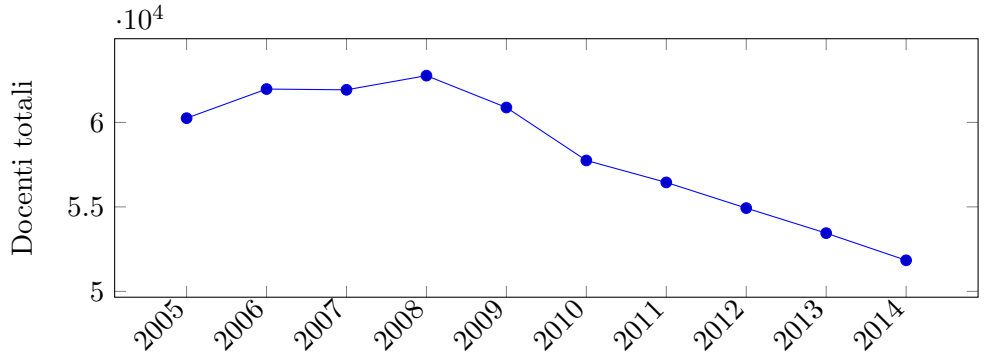


Il grafico a pettine, detto anche a **ordinate disgiunte** o a **bastoncini**, si differenzia dal grafico per punti dall'aggiunta dei segmenti che congiungono i punti del grafico con la loro proiezione su un asse. Affinché un grafico a pettine abbia senso, i segmenti dovranno essere paralleli ad un asse che rappresenta una variabile numerica.

Esempio 1.2.3. **Grafico lineare a tratti** del numero di docenti universitari nel periodo 2005-2014.



(il numero di docenti è scritto in notazione scientifica). A volte, per ottenere una migliore visualizzazione dei dati, il punto di intersezione degli assi non viene fatto coincidere con il valore "zero" della scala. Con questo accorgimento il precedente grafico diventa:



L'effetto visivo è decisamente diverso: mentre nel primo grafico si percepisce una lieve diminuzione, nel secondo si riconosce una vera e propria decimazione. Questo stratagemma è ben noto agli esperti di marketing e propaganda, che lo sfruttano comunemente pro o contro un determinato prodotto, servizio, governo ecc. a seconda della loro convenienza.[4]

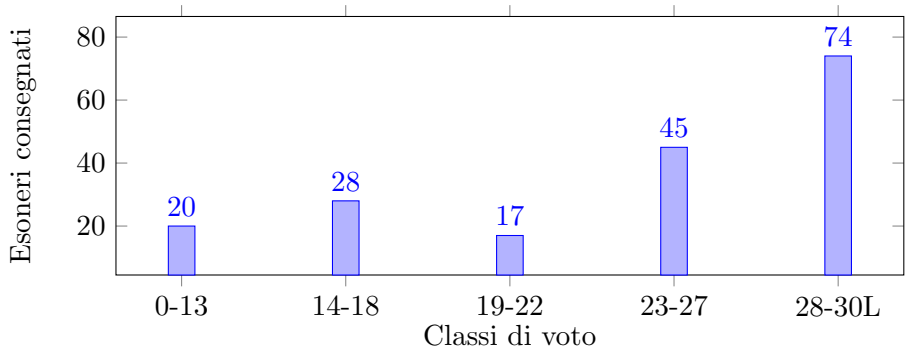
Notiamo che, di norma, un grafico lineare a tratti ha senso quando entrambe le variabili sono di tipo numerico e si ottiene dal grafico per punti congiungendo con dei segmenti le coppie di punti aventi ascisse consecutive.

Quando le coppie di dati sono tante, potrebbe essere preferibile raggruppare i possibili valori delle ordinate e/o delle ascisse in classi omogenee. Anche se i valori tipici per il numero di classi sono tra 5 e 10, la scelta migliore deve essere fatta in maniera soggettiva ed empirica, anche provando varie situazioni fino a trovare quella che porta ai grafici più significativi. Ad esempio, dovendo rappresentare

[4]Per questo ed altri aspetti parastatistici non ci stancheremo mai di raccomandare la lettura del classico testo di Darrell Huff "How to lie with statistic" (1954).

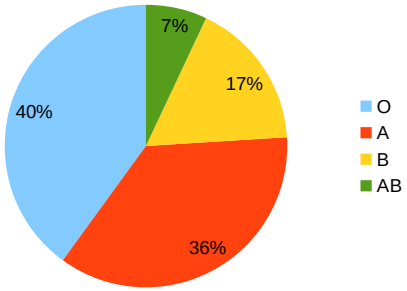
graficamente i risultati di un'esame scritto, può risultare utile dividere lo spettro dei voti (da 0 a 30 e lode) in un numero inferiore di classi.

Esempio 1.2.4. **Grafico a barre** dei risultati del primo esonero di Matematica, divisi per classe di voto. Corso di Laurea in Scienze Naturali, Sapienza Università di Roma, anno accademico 2015-16.



Un grafico a barre, detto anche a **canne d'organo**, non è altro che un diagramma a pettine in cui i segmenti sono sostituiti con rettangoli della stessa base ed altezza variabile.

Esempio 1.2.5. **Aerogramma** con le percentuali di diffusione dei gruppi sanguigni in Italia.



Gli aerogrammi, detti anche **grafici a torta**, si utilizzano quando si vuole mettere in evidenza, più che le misure effettive delle singole grandezze, i loro mutui rapporti. Si costruiscono dividendo un cerchio in tanti settori circolari quante sono le classi da rappresentare, con ciascun settore di ampiezza proporzionale alla consistenza percentuale della classe corrispondente.

Quando la variabile è numerica continua e assume come possibili valori i numeri reali all'interno di un dato intervallo, una efficace rappresentazione consiste nel dividere l'intervallo in n intervalli di uguale ampiezza (come già detto, buoni valori di n sono compresi tra 5 e 10) e contare quanti dati cadono in ciascun intervallo. Ad esempio, la durata dei 26 brani contenuti nell'album "Sandinista!" dei Clash può essere efficacemente illustrata dalla tabella:

Durata (minuti)	numero brani
$1 < t \leq 2$	1
$2 < t \leq 3$	3
$3 < t \leq 4$	8
$4 < t \leq 5$	8
$5 < t \leq 6$	6

Concludiamo la sezione ricordando che i più diffusi pacchetti software di produttività, come ad esempio il gratuito[5] LibreOffice, mettono a disposizione nei loro fogli di calcolo tutti gli strumenti per creare i grafici dei tipi appena menzionati. L'aerogramma sui gruppi sanguigni è stato creato con LibreOffice, gli altri grafici di questa sezione con il pacchetto LaTeX/pgfplots.

1.3. Frequenze, frequenze relative e mode

Da tutti gli esempi fin qui descritti, appare chiaro che ogni indagine e rilevazione statistica è descritta astrattamente da una funzione $X\colon \Omega \to V$ detta **distribuzione di dati** o **dati grezzi**, dove Ω è la popolazione, X è la variabile e V è l'insieme dei valori che la variabile può assumere.

Avviso importante: salvo avviso contrario, assumeremo sempre che Ω e V siano **insiemi finiti**. In particolare, ogni variabile numerica continua viene discretizzata considerando come V un insieme finito di intervalli nella retta reale.

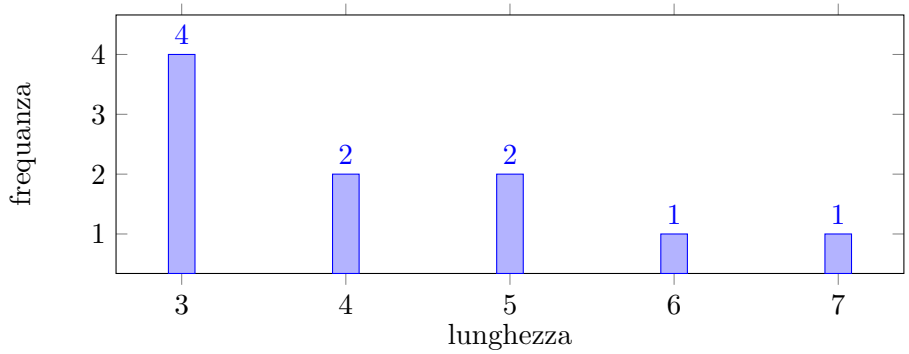
Dopo aver definito chiaramente chi sono Ω, X e V (questo compito spetta alla statistica metodologica, che non è oggetto di queste note), solitamente il primo passo per la comprensione dei dati raccolti consiste nel determinare la **distribuzione delle frequenze**: per ogni possibile valore $x \in V$ si indica quanti sono gli elementi della popolazione Ω in cui la variabile X assume il valore x. I modi più semplici e maggiormente usati per descrivere una distribuzione delle frequenze è mediante tabelle, diagrammi ed istogrammi.

Esempio 1.3.1. Ventuno compagni di scuola festeggiano al ristorante la fine dell'anno scolastico e ciascuno di loro ordina un primo a scelta tra: bucatini all'amatriciana, spaghetti alla carbonara, penne all'arrabbiata e tonnarelli alla gricia. Il cameriere che prende le ordinazioni non scrive ventuno primi, ma si limita a consegnare in cucina una tabella delle frequenze del tipo

amatriciana	10
carbonara	5
arrabbiata	3
gricia	3

Esempio 1.3.2. Dall'analisi della lunghezza delle parole indicanti i numeri naturali da 1 a 10 nella lingua italiana, ne risulta il seguente diagramma delle frequenze:

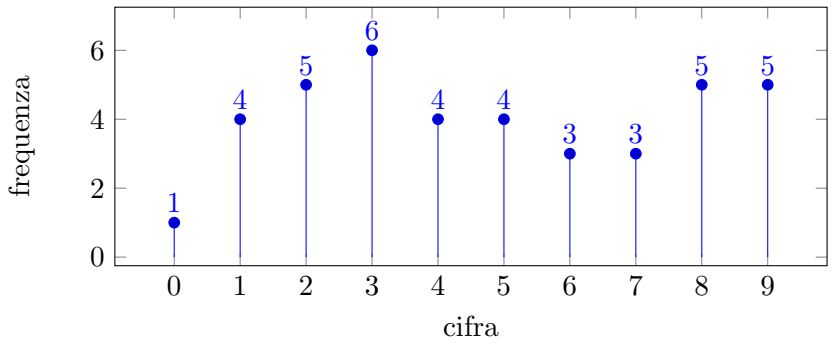
[5] A gennaio 2016.



ESEMPIO 1.3.3. Le prime 40 cifre decimali dopo la virgola del numero $\pi \simeq 3,14...$ sono

$$1415926535897932384626433832795028841971\ .$$

La cifra 0 compare una volta, la cifra 1 compare quattro volte eccetera. La distribuzione delle frequenze associata è descritta dal seguente grafico a pettine[6]



La **frequenza relativa** di un valore è la frequenza diviso la consistenza numerica della popolazione:

$$\text{frequenza relativa di } x = \frac{\text{frequenza di } x}{\text{numero di elementi in } \Omega}, \qquad x \in V\ .$$

È del tutto chiaro che le frequenze relative sono numeri reali compresi tra 0 e 1.

ESEMPIO 1.3.4. Da 150 analisi effettuate su campioni di sangue, 78 sono risultati di tipo 0, 48 di tipo A, 15 di tipo B e 9 di tipo AB. La tabella delle frequenze relative risulta quindi essere:

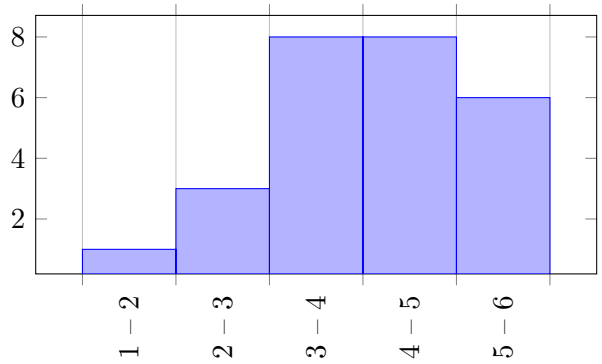
0	A	B	AB
$\frac{78}{150} = 0,52$	$\frac{48}{150} = 0,32$	$\frac{15}{150} = 0,1$	$\frac{9}{150} = 0,06$

[6]Essendo il numero π irrazionale può apparire strana ed inspiegabile la differenza tra il numero di volte in cui compare la cifra 0 ed il numero di volte in cui compare la cifra 3. La risposta è che 40 cifre è un numero troppo basso per non avere fluttuazioni statistiche significative: un'analisi delle prime 10.000.000 cifre mostra che le frequenze sono tutte comprese tra un minimo di 999333 (cifra 1) ed un massimo di 1001093 (cifra 4), vedi http://blogs.sas.com/content/iml/2015/03/12/digits-of-pi.html.

Definizione 1.3.5. La **moda** di una distribuzione di dati è il valore corrispondente alla frequenza più grande. Se vi sono due o più valori aventi frequenza massima si parla di *mode* e distribuzioni **bimodali** o **multimodali**.

Le mode dei precedenti tre esempi sono quindi i bucatini all'amatriciana, la lunghezza 3 ed il gruppo 0.

Quando la variabile è numerica continua che prende valori in intervallo, un modo per rappresentare graficamente le frequenze è mediante un **istogramma**. Un istogramma consiste in un insieme di rettangoli adiacenti, aventi base sull'asse orizzontale; le basi sono gli intervalli che definiscono le classi (i punti medi delle basi sono i valori centrali delle classi). Se le classi hanno tutte la stessa ampiezza le altezze dei rettangoli sono uguali, o proporzionali, alle corrispondenti frequenze assolute (oppure relative o percentuali). Se invece le classi sono di ampiezza diversa, i rettangoli hanno ancora base uguale alla corrispondente ampiezza della classe, e area (non più altezza!) corrispondente alla frequenza: l'altezza del rettangolo sarà uguale, o proporzionale, al rapporto fra la frequenza e l'ampiezza di classe. Tale rapporto si chiama densità di frequenza. In entrambi i casi quindi l'area di ogni rettangolo è uguale, o proporzionale, alla frequenza della classe. Ad esempio, l'istogramma delle frequenze dei brani di "Sandinista!", divise per classi di durata è:



In alcune situazioni può essere più utile presentare i dati nella cosiddetta **forma cumulativa**: la frequenza totale di tutti i valori minori od uguali ad un certo valore viene detta **frequenza cumulativa.** Una tabella che presenti frequenze cumulative è detta tabella di distribuzione cumulativa di frequenza. Si possono cumulare frequenze assolute, relative e percentuali; l'ultimo valore che compare nella tabella sarà uguale al numero totale di dati per le frequenze assolute, uguale a 1 per le frequenze relative e uguale a 100 per quelle percentuali.

Ad esempio, nel precedente esempio dell'album "Sandinista!", la tabella delle frequenze cumulative assolute diventa:

Durata (minuti)	numero brani
$t \leq 2$	1
$t \leq 3$	4
$t \leq 4$	12
$t \leq 5$	20
$t \leq 6$	26

Una distribuzione cumulativa viene rappresentata con un grafico detto **poligono cumulativo** o **ogiva**; il grafico si ottiene riportando sulle ascisse i limiti superiori delle classi e, per ciascuno di essi, in ordinata la frequenza cumulativa della corrispondente classe, e unendo poi tra loro i punti ottenuti.

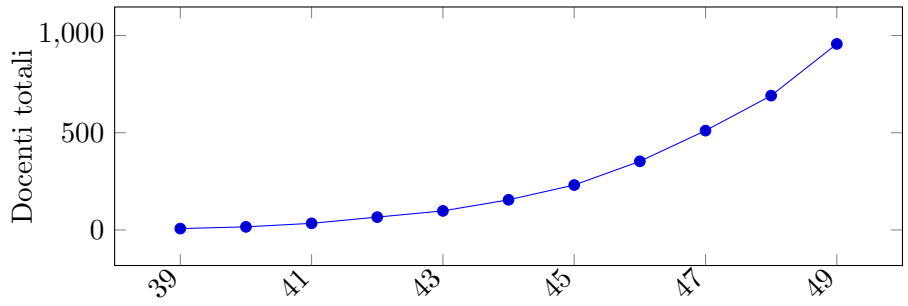
Esempio 1.3.6. La seguente tabella descrive il numero di professori ordinari con meno di 50 anni in servizio al 31 dicembre 2014 nelle università italiane:

Età	39	40	41	42	43	44	45	46	47	48	49
Numero	7	9	18	32	32	57	76	122	158	180	266

La corrispondente tabella cumulativa è pertanto:

Età	≤ 39	≤ 40	≤ 41	≤ 42	≤ 43	≤ 44	≤ 45	≤ 46	≤ 47	≤ 48	≤ 49
N.	7	16	34	66	98	155	231	353	511	691	957

Si noti che i 957 ordinari con meno di 50 anni rappresentavano, al momento della rilevazione, il $7,2\%$ del totale degli ordinari in servizio. Il relativo poligono cumulativo è:



Esercizi.

Esercizio 1.4. Le prime 80 cifre decimali di π dopo la virgola sono

$$14159265358979323846264338327950288419716939937510582097494459230781640628620899 .$$

Disegnare i grafici a pettine delle frequenze e delle frequenze relative e dire se tale distribuzione ha una sola moda oppure se è multimodale.

Esercizio 1.5. Si consideri la distribuzione delle lunghezze delle parole corrispondenti ai numeri interi tra 1 e N nella lingua italiana. Tra tutti gli interi $N > 10$, determinare il più piccolo per cui la distribuzione diventa bimodale.

1.4. Tipi di medie

Arrivati a questo punto abbandoniamo gli aspetti discorsivi e divulgativi della statistica e iniziamo ad occuparci degli aspetti matematici. Prima di iniziare a parlare di dati statistici veri e propri, è utile soffermarci su alcuni aspetti matematici riguardanti il concetto di media.

Per **media aritmetica** (spesso detta semplicemente media) di due numeri x_1, x_2 si intende la metà della loro somma:

$$\bar{x} = \frac{x_1 + x_2}{2}\,.$$

Più in generale la media aritmetica di n numeri $x_1, \dots, x_n$ è uguale a

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i\,,$$

dove Σ è il simbolo di sommatoria: date le quantità $a_1, \dots, a_n$ si pone

$$\sum_{i=1}^{n} a_i = a_1 + \cdots + a_n,$$

ogniqualvolta la somma a destra del segno di uguaglianza è ben definita. Più in generale se $m \leq n$ si pone

$$\sum_{i=m}^{n} a_i = a_m + \cdots + a_n.$$

Esempio 1.4.1. La media aritmetica di $2, 3, 5, 7$ è

$$\frac{2+3+5+7}{4} = \frac{17}{4} = 4,25\,.$$

In ambito statistico, le media aritmetica viene talvolta detta **media campionaria**. La media aritmetica gode di alcune proprietà matematiche elencate nel seguente teorema.

Teorema 1.4.2. *Sia $\bar{x}$ la media aritmetica della successione numerica $x_1, \dots, x_n$. Allora:*

(1) *la media aritmetica della successione $x_1, \dots, x_n, \bar{x}$ è ancora $\bar{x}$;*
(2) *la media aritmetica $\bar{x}$ è il punto di minimo assoluto della funzione*

$$f(t) = (t - x_1)^2 + \cdots + (t - x_n)^2 = \sum_{i=1}^{n} (t - x_i)^2\,,$$

il cui valore minimo assoluto è uguale a

$$f(\bar{x}) = \sum_{i=1}^{n} (\bar{x} - x_i)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\,.$$

Dimostrazione. Tenendo conto che per ipotesi si ha $x_1 + \cdots + x_n = n\bar{x}$, la media aritmetica della successione $x_1, \dots, x_n, \bar{x}$ è uguale a

$$\frac{x_1 + \cdots + x_n + \bar{x}}{n+1} = \frac{n\bar{x} + \bar{x}}{n+1} = \frac{(n+1)\bar{x}}{n+1} = \bar{x}\,.$$

La funzione $f(t)$ è derivabile su tutto $\mathbb{R}$ e la sua derivata è uguale a

$$f'(t) = 2(t - x_1) + \cdots + 2(t - x_n) = 2nt - 2(x_1 + \cdots + x_n) = 2n(t - \bar{x})\,.$$

Quindi $f'(t) < 0$ per $t < \bar{x}$, $f'(t) > 0$ per $t > \bar{x}$ e $f'(t) = 0$ per $t = \bar{x}$; questo è più che sufficiente per dedurre che $t = \bar{x}$ è l'unico punto di minimo assoluto della funzione $f(t)$. Il valore minimo è quindi uguale a

$$\begin{aligned} f(\bar{x}) &= \sum_{i=1}^{n}(\bar{x} - x_i)^2 = \sum_{i=1}^{n}(\bar{x}^2 - 2\bar{x}x_i + x_i^2) \\ &= n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i^2 = n\bar{x}^2 - 2n\bar{x}^2 + \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \,. \end{aligned}$$

□

Quando i numeri $x_1, \ldots, x_n$ *sono tutti positivi* è tavolta utile considerare altri tipi di medie:

(1) la *media quadratica* è la radice quadrata della media aritmetica dei quadrati, ossia
$$m_2 = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}} \,,$$
(2) la *media geometrica* è la radice n-esima del prodotto, ossia
$$m_g = \sqrt[n]{x_1 x_2 \cdots x_n} \,,$$
(3) la *media armonica* è l'inverso della media aritmetica degli inversi, ossia
$$m_h = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \cdots + \frac{1}{x_n}\right)} = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \,.$$

Ad esempio, per la succesione di numeri $x_1 = 1, x_2 = 2, x_3 = 3$ si ha

$$\bar{x} = 2, \quad m_2 = \sqrt{\frac{14}{3}} \simeq 2,16\,, \quad m_g = \sqrt[3]{6} \simeq 1,82\,, \quad m_h = \frac{3}{1 + \frac{1}{2} + \frac{1}{3}} = \frac{18}{11} \simeq 1,63\,.$$

La media armonica compare in pratica più spesso di quanto si tende ad immaginare: consideriamo ad esempio un pilota di auto sportive che effettua il primo giro di pista alla velocità media di 200 Km/h ed il secondo giro alla velocità media di 300 Km/h. La velocià media dei due giri è allora data dalla media armonica delle due (esercizio: perché?) ed è quindi uguale a

$$\frac{2}{\frac{1}{200} + \frac{1}{300}} = 240 \text{ Km/h.}$$

Le relazioni più significative tra le varie medie sono descritte dal seguente teorema.

Teorema 1.4.3. *Siano $\bar{x}, m_2, m_g$ e m_h le medie aritmetica, quadratica, geometrica ed armonica della successione $x_1, \ldots, x_n$ di numeri reali positivi. Allora*

$$m_2 \geq \bar{x} \geq m_g \geq m_h \,.$$

Inoltre, se i numeri x_i non sono tutti uguali tra loro valgono le disuguaglianze strette

$$m_2 > \bar{x} > m_g > m_h \,.$$

DIMOSTRAZIONE. La dimostrazione di $m_2 \geq \bar{x}$ segue immediatamente dal teorema precedente. Infatti, essendo m_2 ed $\bar{x}$ entrambi ≥ 0, si ha $m_2 \geq \bar{x}$ se e solo se $m_2^2 - \bar{x}^2 \geq 0$. Abbiamo visto che

$$m_2^2 - \bar{x}^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n} = \frac{1}{n}\sum_{i=1}^n (\bar{x} - x_i)^2 \geq 0$$

Se gli x_i non sono tutti uguali, allora esiste almeno un indice j tale che $x_i \neq \bar{x}$ e di conseguenza

$$m_2^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^n (\bar{x} - x_i)^2 \geq (\bar{x} - x_j)^2 > 0\,.$$

Per semplicità dimostriamo le disuguaglianze $\bar{x} \geq m_g \geq m_h$ nel caso $n = 2$; il caso generale si può dimostrare in maniera del tutto simile, anche se più complicata. Per quanto riguarda il confronto tra media aritmetica e media geometrica della successione x_1, x_2 si ha:

$$\bar{x}^2 - m_g^2 = \frac{x_1^2 + x_2^2 + 2x_1x_2}{4} - x_1x_2 = \frac{x_1^2 + x_2^2 - 2x_1x_2}{4} = \left(\frac{x_1 - x_2}{2}\right)^2 \geq 0,$$

ed è chiaro che vale $\bar{x}^2 - m_g^2 = 0$ se e solo se $x_1 = x_2$. Per quanto riguarda il confronto tra media armonica e media geometrica si ha:

$$\begin{aligned} m_g^2 - m_h^2 &= x_1x_2 - \frac{4x_1^2x_2^2}{(x_1 + x_2)^2} = \frac{x_1x_2}{(x_1 + x_2)^2}(x_1^2 + x_2^2 - 2x_1x_2) \\ &= \frac{x_1x_2}{(x_1 + x_2)^2}(x_1 - x_2)^2 \geq 0\,. \end{aligned}$$

Anche in questo caso notiamo che se $x_1 \neq x_2$ allora $m_g > m_h$.

□

OSSERVAZIONE 1.4.4. Esiste un altro modo di dimostrare le disuguaglianze $\bar{x} \geq m_g$ e $\bar{x} \geq m_h$ basato sul seguente lemma.

LEMMA 1.4.5. *Sia $f\colon]0, +\infty[\to \mathbb{R}$ una funzione derivabile due volte con derivata seconda negativa, ossia con derivata prima decrescente. Allora per ogni $x_1, \ldots, x_n > 0$ si ha*

$$f(\bar{x}) \geq \frac{f(x_1) + \cdots + f(x_n)}{n}\,.$$

DIMOSTRAZIONE. Senza dimostrazione. □

Ad esempio, la funzione $f(t) = \log(t)$ soddisfa le ipotesi del lemma in quanto

$$\log(t)'' = -\frac{1}{t^2} < 0$$

e dunque si ha

$$\log(\bar{x}) \geq \frac{\log(x_1) + \cdots + \log(x_n)}{n}\,,$$

che equivale a $n\log(\bar{x}) \geq \log(x_1) + \cdots + \log(x_n) = \log(x_1 \cdots x_n)$ e che prendendo gli esponenziali ci fornisce la disuguaglianza

$$\bar{x}^n = e^{n\log(\bar{x})} \geq e^{\log(x_1 \cdots x_n)} = x_1 \cdots x_n\,.$$

Anche la funzione $f(t) = -\frac{1}{t}$ soddisfa le ipotesi del lemma e quindi vale la disuguaglianza

$$-\frac{1}{\bar{x}} \geq \frac{1}{n}\left(-\frac{1}{x_1} - \cdots - \frac{1}{x_n}\right)$$

che equivale alla

$$\frac{n}{\bar{x}} \leq \frac{1}{x_1} + \cdots + \frac{1}{x_n}\,.$$

Esercizi.

Esercizio 1.6. Siano L, M due rette parallele nel piano e siano p, q punti della retta L. Detto r il punto medio del segmento pq, determinare il punto di M per cui la somma delle distanze dai punti p, q, r è minima.

Esercizio 1.7. Tra tutti i rettangoli di perimetro 4, il quadrato di lato 1 è quello di area massima. Giustificare la risposta.

Esercizio 1.8. Durante il suo viaggio da Furore a Povo, Gianni effettua tre rifornimenti di carburante. Nel primo acquista 30 litri di Gasolio[7] a 1,32 Euro/Litro, nel secondo acquista 30 litri a 1,4 Euro/Litro e nel terzo 30 litri a 1,3 Euro/Litro. Determinare il prezzo medio di acquisto.

Esercizio 1.9. Durante il suo viaggio da Furore a Povo, Pinotto effettua tre rifornimenti di carburante. Nel primo acquista 40 euro di Gasolio a 1,32 Euro/Litro, nel secondo acquista 40 euro a 1,4 Euro/Litro e nel terzo 40 euro a 1,3 Euro/Litro. Determinare il prezzo medio di acquisto.

Esercizio 1.10. Tizio e Caio viaggiano molto in auto per motivi di lavoro ed ogni giorno si riforniscono al medesimo distributore. Ogni giorno Tizio si rifornisce con 40 litri di benzina, mentre Caio si rifornisce con 50 euro di benzina. Chi tra Tizio e Caio paga mediamente meno un litro di benzina?

1.5. Mediane, quantili, quartili e percentili

Supponiamo di aver misurato l'altezza di tutti gli studenti maschi della Sapienza; nella sezione precedente abbiamo definito l'altezza media come la somma di tutte le altezze diviso il numero di studenti. In questa sezione definiamo l'altezza mediana come quella che, detto molto informalmente, separa la metà degli "alti" dalla metà dei "bassi". Se, per ipotesi, l'altezza mediana fosse di 175 centimetri, significherebbe chè metà studenti sono più alti di 175 centimetri e metà sono più bassi di 175 cm.

Guardando agli aspetti matematici, non sempre è possibile dividere una popolazione a metà, e non sempre esiste un unico valore che separa in due parti uguali. Ecco quindi che la nozione di mediana viene formalizzata nel modo seguente:

Definizione 1.5.1. Sia

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n\,.$$

una successione non decrescente di n numeri reali. La **mediana** M di tale successione è uguale a:

(1) $M = x_{\frac{n+1}{2}}$ se n è dispari;

[7] I prezzi indicati sono in linea con quelli in vigore a gennaio 2015.

(2) $M = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+2}{2}}\right)$ se n è pari.

In altri termini, se n è dispari si prende il valore di mezzo,

$$3,\ \mathbf{5},\ 8 \qquad M = 5\,,$$

$$1,\ 2,\ \mathbf{4},\ 7,\ 9 \qquad M = 4\,,$$

mentre se n è pari si prende la media aritmetica dei due valori di mezzo

$$3,\ \mathbf{4},\ \mathbf{5},\ 9 \qquad M = \frac{4+5}{2} = 4,5\,,$$

$$2,\ 3,\ \mathbf{4},\ \mathbf{4},\ 6,\ 9 \qquad M = \frac{4+4}{2} = 4\,.$$

Naturalmente si può definire anche la mediana di una qualsiasi successione finita di numeri: a tal fine basta ordinare i numeri in ordine crescente per ricondursi alla definizione precedente.

Esempio 1.5.2. La mediana della successione

$$2,\ 7,\ 4,\ 7,\ 8,\ 2,\ 3$$

è uguale a 4, essendo tale il valore mediano della medesima successione riordinata in maniera crescente:

$$2,\ 2,\ 3,\ \mathbf{4},\ 7,\ 7,\ 8\,.$$

Data una successione di n numeri, abbiamo quindi il seguente metodo di calcolo della mediana M:

(1) se $\frac{n}{2}$ è un numero frazionario e k è l'arrotondamento di $\frac{n}{2}$ all'intero superiore, ossia $\frac{n}{2} < k < \frac{n}{2} + 1$, allora M è uguale al k-esimo elemento della successione ordinata in ordine crescente.

(2) se $\frac{n}{2} = k$ è un numero intero, allora la mediana è uguale alla media aritmetica dei termini alle posizioni k e $k+1$ rispetto all'ordinamento crescente.

Media e mediana sono detti **indici di posizione** o **indici di tendenza centrale**, perché descrivono attorno a quale valore è centrato l'insieme di dati.

Si preferisce usare la mediana in quelle situazioni dove non interessano tanto i valori numerici delle grandezze in esame, quanto piuttosto il loro ordinamento. La mediana è preferibile alla media quando si vogliono eliminare gli effetti di valori estremi molto diversi dagli altri dati: la ragione è che la mediana non utilizza tutti i dati, ma solo il dato centrale o i due dati centrali. Tuttavia occorre mettere in evidenza che l'utilizzare solo i dati centrali rende la mediana poco sensibile a tutti gli altri valori dei dati e questo può costituire un limite di questo indice. L'uso della mediana come indice per descrivere le caratteristiche dei dati ha inoltre lo svantaggio di dover prima riordinare i dati in ordine crescente, il che non è richiesto per il calcolo della media.

Esempio 1.5.3. Un'impresa si pubblicizza nei suoi annunci di lavoro dicendo che il guadagno medio dei propri ingegneri è di 40.000 euro all'anno. Da un'analisi più approfondita si scopre che dei 5 ingegneri presenti, i primi quattro guadagnano 25.000 euro l'anno, ed il figlio del proprietario, anch'esso ingegnere, guadagna 100.000 euro

annui. In questo caso l'indicatore mediano di 25.000 euro/anno risulterebbe più adeguato per descrivere la reale situazione.

Tra le proprietà della mediana c'è quella di minimizzare lo scarto assoluto di una distribuzione numerica, nel senso descritto dal seguente teorema.

TEOREMA 1.5.4. *La mediana di una successione di numeri $x_1 \leq \cdots \leq x_n$ è un punto di minimo assoluto della funzione*

$$f(y) = \sum_{i=1}^{m} |y - x_i| \, .$$

DIMOSTRAZIONE. Denotando con y_0 la mediana, bisogna dimostrare che per ogni $d > 0$ si hanno le disuguaglianze $f(y_0 + d) \geq f(y_0)$ e $f(y_0 - d) \geq f(y_0)$; dimostriamo solamente la prima, essendo la dimostrazione della seconda del tutto simile. Sia $k \leq n$ il più grande indice tale che $x_k \leq y_0$; per definizione di mediana si ha $2k \geq n$. Se $1 \leq i \leq k$ si ha quindi $|y_0 + d - x_i| = |y_0 - x_i| + d$, mentre se $k < i \leq n$, dalla disuguaglianza triangolare $|a - b| \geq |a| - |b|$, si ha

$$|y_0 + d - x_i| = |x_i - y_0 - d| \geq |x_i - y_0| - d \, .$$

Dunque

$$\begin{aligned} f(y_0 + d) &= \sum_{i=1}^{m} |y_0 + d - x_i| = \sum_{i=1}^{k} |y_0 + d - x_i| + \sum_{i=k+1}^{m} |y_0 + d - x_i| \\ &\geq \sum_{i=1}^{k} (|y_0 - x_i| + d) + \sum_{i=k+1}^{m} (|y_0 - x_i| - d) = f(y_0) + (2k - n)d \geq f(y_0) \, . \end{aligned}$$

Si noti che in generale la mediana non è l'unico punto di minimo assoluto; ad esempio, nel caso $n = 2$, ogni numero compreso tra x_1 e x_2 è un punto di minimo assoluto. □

Oltre alla mediana, che divide a metà un insieme di dati ordinati, si possono definire altri indici di posizione, detti **quantili**, che dividono l'insieme di dati ordinati in un dato numero di parti uguali. Questi indici di posizione non centrale sono usati soprattutto per ampi insiemi di dati. I **quartili** sono un caso particolare dei quantili, e si ottengono dividendo l'insieme di dati ordinati in quattro parti uguali.

DEFINIZIONE 1.5.5. Sia q un numero reale strettamente compreso tra 0 e 1. Il termine P_q di una successione

$$x_1 \leq x_2 \leq \cdots \leq x_{n-1} \leq x_n \, .$$

non decrescente di n numeri reali si definisce nel modo seguente:

(1) se qn è un numero frazionario, si pone $P_q = x_k$, dove k è l'arrotondamento di qn all'intero superiore, ossia $qn < k < qn + 1$;

(2) se $qn = k$ è un numero intero, si definisce P_q uguale alla media aritmetica di x_k e x_{k+1}.

Detto in maniera molto imprecisa ma comprensibile anche ai politicanti, il termine P_q di una successione ordinata è quello che separa i minori nq elementi dai maggiori $n(1 - q)$ elementi.

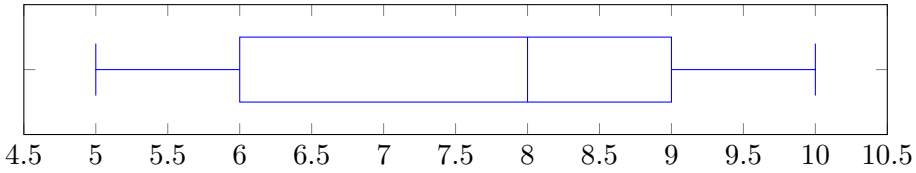


FIGURA 1.1. Boxplot senza dati anomali di una distribuzione con minimo $\min = 5$, primo quartile $Q_1 = 6$, mediana $M = 8$, terzo quartile $Q_3 = 9$ e massimo $\max = 10$.

Dunque $P_{0,5}$ coincide con la mediana. I valori $Q_1 = P_{0,25}$ e $Q_3 = P_{0,75}$ sono detti rispettivamente **primo quartile** e **terzo quartile**. Similmente, per ogni intero k compreso tra 1 e 99 il termine $P_{\frac{k}{100}}$ viene detto **k-esimo percentile**. Similmente, il terzo decile è $P_{0,3}$, il secondo terzile è $P_{2/3}$, il quinto sestile è $P_{5/6}$ e così via.

ESEMPIO 1.5.6. Calcoliamo massimo, minimo, mediana e quartili della distribuzione di numeri reali

$$1,\ 4,\ 2,\ 5,\ 6,\ 3,\ 2,\ 3,\ 3,\ 5,\ 1\,.$$

Dopo aver scritto gli 11 numeri in ordine crescente troviamo

$$1,\ 1,\ 2,\ 2,\ 3,\ 3,\ 3,\ 4,\ 5,\ 5,\ 6\,.$$

e quindi minimo e massimo sono 1 e 6 rispettivamente. Poiché

$$\frac{25\cdot 11}{100} = 2,75, \qquad \frac{50\cdot 11}{100} = 5,5, \qquad \frac{75\cdot 11}{100} = 8,25,$$

ne segue che primo quartile, mediana e terzo quartile sono il terzo, il sesto ed il nono termine della successione ordinata, ossia $Q_1 = 2, M = 3, Q_3 = 5$.

Una figura utile a visualizzare mediane e quartili di una distribuzione di dati numerici è il **boxplot**. Un boxplot si ottiene disegnando un rettangolo (il box) con i lati verticali corrispondenti al primo e terzo quartile, tagliato da un'altra linea verticale in corrispondenza della mediana. Al rettangolo vengono poi aggiunti due "baffi" che indicano il massimo ed il minimo dei dati "non anomali" della distribuzione (Figura 1.1). Per dato anomalo (in inglese outlier) si intende un dato molto più grande o molto più piccolo della maggioranza dei dati; la loro presenza viene indicata sul boxplot con dei pallini sulla retta orizzontale che divide il rettangolo a metà.

OSSERVAZIONE 1.5.7. Altri testi ed alcuni tipi di software usano definizioni più complesse di quartili e percentili che fanno uso di interpolazioni lineari. Di conseguenza i valori risultano diversi da quelli definiti precedentemente, anche se in pratica le differenze risultano poco significative per grandi valori di n.

La seguente tabella illustra l'algoritmo usato da LibreOffice per il calcolo dei quartili di una successione $x_1 \le \cdots \le x_n$; ci sono 4 casi da trattare, a seconda del resto della divisione di n per 4:

	Primo quartile	Terzo quartile
$n = 4k-3$	x_k	x_{n-k+1}
$n = 4k-2$	$\frac{3}{4}x_k + \frac{1}{4}x_{k+1}$	$\frac{3}{4}x_{n-k+1} + \frac{1}{4}x_{n-k}$
$n = 4k-1$	$\frac{1}{2}x_k + \frac{1}{2}x_{k+1}$	$\frac{1}{2}x_{n-k+1} + \frac{1}{2}x_{n-k}$
$n = 4k$	$\frac{1}{4}x_k + \frac{3}{4}x_{k+1}$	$\frac{1}{4}x_{n-k+1} + \frac{3}{4}x_{n-k}$

Si noti che il calcolo dei quartili mediante interpolazione lineare coincide con la precedente definizione quando $n = 4k-3$. Segnaliamo che questa non è l'unica alternativa possibile e che esistono altri metodi di calcolo dei quartili.

Esercizi.

Esercizio 1.11. Scrivete:

(1) 5 numeri distinti tali che la media aritmetica coincida con la mediana;
(2) 5 numeri distinti tali che la media aritmetica sia minore della mediana;
(3) 5 numeri distinti tali che la media aritmetica sia maggiore della mediana.

Esercizio 1.12. Per ogni intero n compreso tra 5 e 10 scrivete una successione di n numeri il cui boxplot sia quello della Figura 1.1.

Esercizio 1.13. La Tabella 1 mostra la classifica finale del campionato di calcio di Serie A 2014-15. Per ciascuna delle sette variabili numeriche riportate, calcolare la mediana, i quartili e disegnare il corrispondente boxplot.

Esercizio 1.14. È ben noto che, per un accordo tra le aziende costruttrici, le vecchie lampadine ad incandescenza venivano costruite in modo da avere una vita media di circa 1000 ore, nonostante la medesima tecnologia (filamenti di tungsteno sottovuoto) consentisse durate almeno doppie o triple. Dire se una lampadina molto difettosa (in peggio o in meglio) in un campione di 20 lampadine, influisce di più sulla vita media o sulla vita mediana.

Esercizio 1.15. Sia $x\colon \Omega \to V \subset \mathbb{R}$ una distribuzione di dati numerici su una popolazione Ω di n osservazioni. Per ogni $x \in V$, supponiamo che $x = x_\alpha$ per qualche $\alpha \in \Omega$ e indichiamo con:

(1) $s(x)$ = numero di osservazioni $\omega \in \Omega$ tali che $x_\omega \geq x$;
(2) $i(x)$ = numero di osservazioni $\omega \in \Omega$ tali che $x_\omega \leq x$.

Dimostrare che, dato un qualsiasi numero reale $k \in [0,1]$, si verifica una delle seguenti possibilità:

(1) esiste un unico valore $x \in V$ tale che
$$i(x) \geq kn, \qquad s(x) \geq (1-k)n\ ;$$
(2) kn è un intero ed esistono due valori $x, y \in V$ tali che
$$x < y, \qquad i(x) \geq kn, \qquad s(y) \geq (1-k)n\ .$$

Se $k = q/100$, che ralazione esiste tra i valori x, y delle precedenti possiblità ed il q-esimo percentile?

1.6. Indici di dispersione

Gli indici di posizione non tengono conto della variabilità esistente fra i dati; vi sono distribuzioni che, pur avendo la stessa media e la stessa mediana sono molto diverse fra loro. Ad esempio consideriamo quattro amici maschi e quattro amiche femmine che si incontrano in birreria. Nel corso della serata i quattro ragazzi bevono 2 birre a testa, mentre le quattro ragazze bevono nell'ordine 0,1,3 e 4 birre. Dunque le due distribuzioni sono molto diverse pur avendo la stessa media e la stessa mediana.

Per avere una misura grossolana della variabilità di una distribuzine di dati numerici vengono introdotti alcuni **indici di dispersione**. I più noti sono il **campo di variazione**, lo **scarto interquartile**, la **varianza** e la **deviazione standard**.

Definizione 1.6.1. Il **campo di variazione** R di una distribuzione numerica è la differenza tra il valore massimo ed il valore minimo. Lo **scarto interquartile** IQR è la differenza tra il terzo quartile ed il primo quartile.

Ad esempio nella distribuzione $2, 2, 2, 2$ il campo di variazione e lo scarto interquartile sono entrambi uguali a 0, mentre nelle distribuzione $0, 1, 3, 4$ il campo di variazione è $R = 4 - 0 = 4$ e lo scarto interquartile è

$$IQR = Q_3 - Q_1 = \frac{3+4}{2} - \frac{0+1}{2} = 3\,.$$

A volte il campo di variazione e lo scarto interquatile vengono chiamati rispettivamente intervallo di variazione e distanza interquartile. I simbolo R e IQR derivano dai corrispondenti termini inglesi *range* e *interquartile range*.

Lo scarto interquartile è una misura di variabilità significativa nelle stesse condizioni in cui la mediana è preferibile alla media; rispetto al campo di variazione lo scarto interquartile ha il vantaggio di essere poco sensibile all'esistenza di pochi valori anomali. Quando invece si preferisce la media alla mediana, allora l'indice di dispersione più significativo diventa la deviazione standard.

Definizione 1.6.2. Data una distribuzione numerica $x_1, \ldots, x_n$, con $n > 1$ e valor medio $\bar{x}$ si definisce:

(1) la **varianza campionaria**, o più semplicemente **varianza**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\,,$$

(2) la **deviazione standard** come la radice quadrata della varianza campionaria:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ ,$$

Ad esempio la variazione standard di una distribuzione costante è uguale a 0, mentre la deviazione standard delle distribuzione $0, 1, 3, 4$ è uguale a

$$s = \sqrt{\frac{1}{4-1}\left((0-2)^2 + (1-2)^2 + (3-2)^2 + (4-2)^2\right)} = \sqrt{\frac{10}{3}}\,.$$

Per le formula viste nel Teorema 1.4.2, possiamo calcolare varianza e deviazione standard mediante le formule

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}, \qquad s = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}\ . \tag{1.6.1}$$

Si ha $s \geq 0$ e vale $s = 0$ se e solo se la distribuzione è costante.

Osservazione 1.6.3. Per n abbastanza grande la varianza è molto prossima alla media aritmetica dei quadrati delle distanze dal valor medio $\frac{1}{n}\sum(x_i - \bar{x})^2$. I motivi per cui nella definizione di s^2 (e quindi di s) si preferisce dividere per $n-1$ anziché per n riguardano questioni molto tecniche di statistica inferenziale che vanno al di là degli obiettivi di queste note.

Osservazione 1.6.4. I valori di s e s^2 dipendono dall'unità di misura dei dati ed in particolare la deviazione standard misura la dispersione dei dati con la stessa unità di misura della media dei dati, cosa che non accade per la varianza; questa è la ragione principale per cui la deviazione standard è più usata della varianza.

Esempio 1.6.5. La media $\bar{x}$ e la deviazione standard s della distribuzione

$$0,\ 1,\ 3,\ 2,\ 5,\ 6,\ 2,\ 5,$$

sono

$$\begin{aligned}
\bar{x} &= \frac{0+1+3+2+5+6+2+5}{8} = \frac{24}{8} = 3\,,\\
7s^2 &= (0-3)^2 + (1-3)^2 + (3-3)^2 + (2-3)^2\\
&\quad + (5-3)^2 + (6-3)^2 + (2-3)^2 + (5-3)^2 = 32,\\
s &= \sqrt{\frac{32}{7}} \simeq 2,138\ .
\end{aligned}$$

Osservazione 1.6.6. Notiamo che sia la media che la deviazione standard di una distribuzione di dati $\Omega \to V$ possono essere calcolati a partire dalla corrispondente distribuzione di frequenze. Infatti, se $V = \{v_1, \ldots, v_n\}$ ed ogni v_i ha frequenza (assoluta) f_i, allora la somma $\sum_{i=1}^n f_i$ è uguale al numero di unità osservative (gli elementi di Ω) e la media è uguale a

$$\bar{x} = \frac{\sum_{i=1}^n f_i v_i}{\sum_{i=1}^n f_i}\ .$$

Similmente la varianza è uguale a

$$s^2 = \frac{\sum_{i=1}^n f_i (v_i - \bar{x})^2}{(\sum_{i=1}^n f_i) - 1}\ .$$

Queste formule sono ottenute semplicemente raggruppando, nella definizione di $\bar{x}$ ed s^2 gli addendi di ugual valore.

Esempio 1.6.7. Uno studente ha superato l'esame di matematica sublunare (12 CFU) con la votazione di 30 e l'esame di matematica iperuranica (9 CFU) con la votazione di 26. Considerando il CFU come unità osservativa, è come se avesse superato 12 esami da 1 CFU con voto 30 e 9 esami da 1 CFU con voto 26. Pertanto la media e la variazione standard risultano essere

$$\bar{x} = \frac{12\cdot 30 + 9\cdot 26}{21} \simeq 28,3, \qquad s = \sqrt{\frac{12(30-28,3)^2 + 9(26-28,3)^2}{20}} \simeq 2,03\ .$$

Esempio 1.6.8. Ad un campione di 50 donne maggiorenni viene chiesto quanti figli hanno e le risposte vengono riassunte nella seguente tabella di frequenze:

Numero figli	0	1	2	3	4	5
Donne	14	21	12	2	1	0

Il numero mediano di figli è $M = 1$, poiché $14 < 25 < 14 + 21$, il numero medio di figli è

$$\bar{x} = \frac{21 + 24 + 6 + 4}{50} = \frac{55}{50} = 1,1,$$

mentre la deviazione standard è

$$s = \sqrt{\frac{14(0-1,1)^2 + 21(1-1,1)^2 + 12(2-1,1)^2 + 2(3-1,1)^2 + (4-1,1)^2}{49}} \simeq 0,931\,.$$

Osservazione 1.6.9. Il calcolo degli indici di posizione e dispersione è facilissimo per chi dispone di un personal computer con Libreoffice installato: aprendo con tale programma un foglio di calcolo (spreadsheet) in cui i dati numerici $x_1, \dots, x_n$ vengono inseriti sulle righe da 1 a n della colonna A, si hanno le seguenti funzioni:

- media $\bar{x} =$ AVERAGE($A$1:An),
- mediana $M =$ MEDIAN($A$1:An),
- primo quartile $Q_1 =$ QUARTILE($A$1:An;1), cf. Osservazione 1.5.7,
- terzo quartile $Q_3 =$ QUARTILE($A$1:An;3),
- varianza $s^2 =$ VAR($A$1:An),
- deviazione standard $s =$ STDEV($A$1:An).

1.7. Casi atipici e disuguaglianze di Chebishev

A volte le analisi statistiche vengono fatte per individuare i "casi atipici", ossia gli elementi della popolazione con una variabile numerica sufficientemente distante dalla media. Essendo la deviazione standard una quantità omogenea alla media, in quanto espressa nella stessa unità di misura, il modo corretto per impostare il precedente problema è quello di cercare i dati che differiscono dalla media per un certo numero fissato a priori di deviazioni standard.

Definizione 1.7.1 (Casi atipici). Sia $x_1, \dots, x_n$ una distribuzione di dati numerica con media $\bar{x}$ e deviazione standard s. Per ogni numero reale positivo k indichiamo con:

(1) t_k il numero di dati x_i che superano la media di almeno k deviazioni standard, ossia tali che $x_i - \bar{x} \geq ks$;
(2) d_k il numero di dati x_i che sono superati dalla media di almeno k deviazioni standard, ossia tali che $\bar{x} - x_i \geq ks$.

La precedente definizione ha senso per ogni numero reale $k > 0$, tuttavia vengono considerati nella pratica solo i casi relativi a multipli interi della deviazione standard, ossia per $k = 1, 2, 3, \dots$. Per la serie di dati dell'Esempio 1.6.5 si ha $d_1 = t_1 = 1$, $d_2 = t_2 = 0$.

Un limite superiore al numero di casi atipici è fornito dalle disuguaglianze di Chebyshev.

TEOREMA 1.7.2 (Disuguaglianze di Chebyshev). *Sia $x_1, \dots, x_n$ una distribuzione di dati numerici con media $\bar{x}$ e deviazione standard $s > 0$. Nelle notazioni della Definizione 1.7.1, per ogni numero reale positivo $k > 0$ si hanno le disuguaglianze:*

$$t_k + d_k \leq \frac{n-1}{k^2} < \frac{n}{k^2}\,,$$
$$t_k \leq \frac{n-1}{k^2 + \dfrac{n-1}{n}} < \frac{n}{k^2+1}\,,$$
$$d_k \leq \frac{n-1}{k^2 + \dfrac{n-1}{n}} < \frac{n}{k^2+1}\,.$$

Nota: per la validità delle disuguaglianze di Chebishev è fondamentale che $s > 0$, ossia che la distribuzione non sia costante.

DIMOSTRAZIONE. Per dimostrare la disuguaglianza $k^2(t_k + d_k) \leq n-1$, a meno di scambiare l'ordine dei dati, non è restrittivo supporre che $|x_i - \bar{x}| \geq ks$ per ogni $i = 1, \dots, t_k + d_k$. Allora si ha

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \sum_{i=1}^{t_k+d_k}(x_i - \bar{x})^2 \geq (t_k + d_k)k^2 s^2,$$

e dividendo per s^2 troviamo $n - 1 \geq (t_k + d_k)k^2$.

Per la seconda disuguaglianza, indichiamo $y_i = x_i - \bar{x}$ e come sopra possiamo supporre che $y_i = x_i - \bar{x} \geq ks$ per ogni $i = 1, \dots, t_k$. Per ogni numero reale non negativo $b \geq 0$ abbiamo allora

$$\sum_{i=1}^{n}(y_i + bs)^2 \geq \sum_{i=1}^{t_k}(y_i + bs)^2 \geq t_k(ks + bs)^2\,,$$

Siccome $\sum_i y_i = 0$ e $\sum_i y_i^2 = (n-1)s^2$, si ha

$$\sum_{i=1}^{n}(y_i + bs)^2 = (n-1)s^2 + b^2 s^2$$

e quindi, dividendo per s^2 abbiamo dimostrato che per ogni $b \geq 0$ vale la disuguaglianza

$$(n-1) + nb^2 \geq t_k(k+b)^2\,.$$

Ponendo in particolare $b = \dfrac{n-1}{nk}$ si ottiene

$$(n-1) + \frac{(n-1)^2}{nk^2} \geq t_k\left(k + \frac{n-1}{nk}\right)^2$$

che, moltiplicando per k^2 diventa

$$(n-1)\left(k^2 + \frac{n-1}{n}\right) \geq t_k\left(k^2 + \frac{n-1}{n}\right)^2$$

e poi dividendo per $(k^2 + \frac{n-1}{n})^2$

$$t_k \leq \frac{n-1}{k^2 + \dfrac{n-1}{n}} = \frac{n}{k^2\dfrac{n}{n-1} + 1} < \frac{n}{k^2+1}\,.$$

L'ultima disugaglianza segue immediatamente dalla seconda applicata alla distribuzione numerica $-x_1, \ldots, -x_n$. □

Osservazione 1.7.3. Esiste una regola empirica secondo la quale nelle distribuzioni relative alla misurazione di un fenomeno del mondo reale, in cui la variabilità delle misure è data da una molteplicità di fattori casuali, si ha che:

(1) circa il 68% dei dati cade nell'intervallo $\bar{x} \pm s$;
(2) circa il 95% dei dati cade nell'intervallo $\bar{x} \pm 2s$;
(3) circa il $99,5\%$ dei dati cade nell'intervallo $\bar{x} \pm 3s$.

Naturalmente questa regola empirica non vale per distribuzioni costruite artificiosamente.

Per confrontare la variazione di più campioni di dati, ciascuno con media diversa, o misurati in unità di misura diverse, pu essere utile usare una misura di variazione relativa, anziché una misura assoluta come la deviazione standard.

Definizione 1.7.4. Il **coefficiente di variazione esprime** la deviazione standard come percentuale della media:

$$CV = \frac{s}{\bar{x}} \times 100 \, .$$

Il coefficiente di variazione ha senso solo quando $\bar{x} > 0$ ed è indipendente dall'unità di misura usata, poiché la media e la deviazione standard sono espressi nella stessa unità di misura.

Esempio 1.7.5. Sia dato un campione di sacchetti di biscotti. Si assuma che il peso medio sia 400 grammi e che la deviazione standard del peso sia 25 grammi. Si assuma inoltre che il numero medio di biscotti sia 75 e che la deviazione standard del numero di biscotti sia 7. I coefficentidi variazione del peso e del numero di biscotti sono quindi

$$CV(\text{peso}) = \frac{25}{400} \times 100 = 6,25, \qquad CV(\text{biscotti}) = \frac{7}{75} \times 100 = 9,33 \, .$$

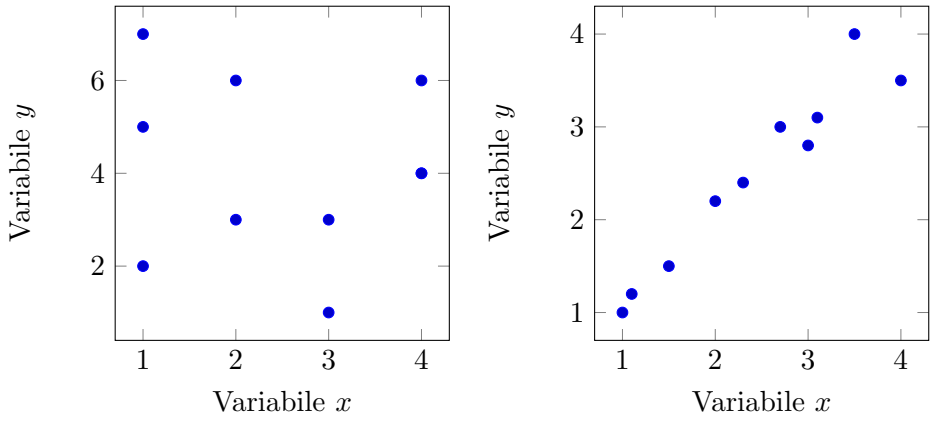
Pertanto, rispetto alla media, il numero di biscotti è più variabile del peso.

1.8. Correlazione statistica

In certe situazioni si eseguono indagini statistiche nelle quali si osservano più variabili su una medesima popolazione. In tal caso, un problema tipico consiste nel chiedersi se esiste una correlazione fra le variabili osservate.

Ad esempio, se la popolazione di riferimento sono gli studenti della Sapienza, ci si può chiedere se esiste una correlazione tra il voto conseguito al test di autovalutazione ed il numero di CFU conseguiti nel primo anno di corso.

Quando si osservano due variabili x e y di tipo numerico, il primo passo utile per indagare qualitativamente l'eventuale dipendenza consiste nel disegnare un grafico, detto **diagramma di dispersione** o **scatterplot**. Si pongono in ascissa i dati relativi a una delle due variabili, in ordinata quelli relativi all'altra variabile e si rappresentano con punti o cerchietti le singole osservazioni. Se esiste una relazione semplice fra le due variabili, il diagramma dovrebbe evidenziarla. Si osservino ad esempio i due diagrammi seguenti, entrambi relativi a popolazioni di 10 unità osservative:



Il primo diagramma non suggerisce che vi sia una correlazione fra le due variabili: i punti sono sparsi apparentemente senza una logica. Il secondo diagramma evidenzia invece una certa regolarità: punti con ascissa piccola hanno ordinata piccola e punti con ascissa grande hanno ordinata grande; in questo caso si dice che esiste una **correlazione diretta** fra le due variabili. Analogamente si parla di **correlazione inversa** se al crescere di una variabile l'altra decresce. Nel secondo diagramma si può inoltre ipotizzare una correlazione tra le due variabili di tipo lineare, nel senso che si i punti si addensano intorno ad una retta, detta retta di regressione.

Per trattare il problema in maniera quantitativa, introdurremo il coefficiente di correlazione r di due variabili x e y, che è un numero reale compreso tra -1 e 1. *Quando il coefficiente r è vicino allo 0 non vi è correlazione, quando r è vicino ad 1 vi è una correlazione diretta, quando r è vicino a -1 vi è una correlazione inversa.*

Definizione 1.8.1. Date n osservazioni congiunte (x_1, y_1), (x_2, y_2),..., (x_n, y_n) di due variabili x e y, si dice **covarianza** delle due variabili x, y il numero

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

dove $\bar{x}$ e $\bar{y}$ sono le medie delle variabili x e y. Il numero

$$r = \frac{s_{xy}}{s_x\, s_y}$$

viene detto **coefficiente di correlazione**, dove s_x e s_y sono le deviazioni standard delle variabili x e y.

Ovviamente il coefficiente di correlazione è definito quando $s_x > 0$ $s_y > 0$, ossia se entrambe le variabili x e y non sono costanti sulla popolazione.

Teorema 1.8.2. *Il coefficiente di correlazione è compreso tra -1 ed 1.*

Dimostrazione. Dimostriamo preliminarmente la cosiddetta *disuguaglianza di Cauchy-Schwarz*: date n coppie di numeri reali (a_1, b_1), (a_2, b_2),..., (a_n, b_n) si ha:

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right) \cdot \left(\sum_{i=1}^{n} b_i^2\right).$$

Denotando

$$p = \sum_{i=1}^{n} a_i b_i, \quad q = \sum_{i=1}^{n} a_i^2, \quad r = \sum_{i=1}^{n} b_i^2$$

vogliamo dimostrare che $qr - p^2 \geq 0$. Se $q = 0$ allora $a_1 = \cdots = a_n = 0$, dunque anche $p = 0$ e la disuguaglianza è vera. Se invece $q > 0$ si ha

$$\sum_{i=1}^{n} (b_i q - a_i p)^2 \geq 0$$

essendo somma di quadrati. Espandendo i quadrati dei binomi nella sommatoria si ottiene

$$\begin{aligned} \sum_{i=1}^{n} (b_i q - a_i p)^2 &= \sum_{i=1}^{n} (b_i^2 q^2 + a_i^2 p^2 - 2 a_i b_i p q) \\ &= r q^2 + q p^2 - 2 p^2 q = r q^2 - p^2 q \geq 0 \end{aligned}$$

e dividendo per il numero positivo q si ottiene $qr - p^2 \geq 0$.

Passiamo adesso alla dimostrazione del teorema. Affermare che $-1 \leq r \leq 1$ equivale a dire che $r^2 \leq 1$, ossia che $s_{xy}^2 \leq s_x^2 s_y^2$. Dato che

$$\begin{aligned} s_{xy}^2 &= \frac{1}{(n-1)^2} \left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right)^2, \\ s_x^2 &= \frac{1}{n-1} \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right), \\ s_y^2 &= \frac{1}{n-1} \left(\sum_{i=1}^{n} (y_i - \bar{y})^2 \right), \end{aligned}$$

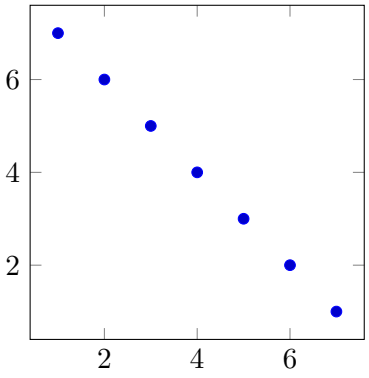
la conclusione segue dalla disuguaglianza di Cauchy-Schwarz, dove $a_i = x_i - \bar{x}$ e $b_i = y_i - \bar{y}$. □

Vediamo adesso tre esempi, con coefficienti di correlazione $-1, 0$ e 1 rispettivamente:

Esempio 1.8.3. Su una popolazione di 7 unità, si considerino le coppie di osservazioni

$$(1,7),\ (2,6),\ (3,5),\ (4,4),\ (5,3),\ (6,2),\ (7,1)\,.$$

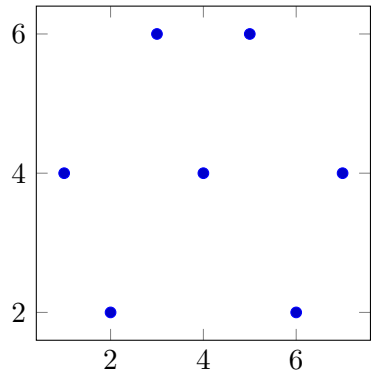
La covarianza è uguale a $s_{xy} = -4,66$, il coefficiente di correlazione è $r = -1$ ed il diagramma di dispersione è



Esempio 1.8.4. Su una popolazione di 7 unità, si considerino le coppie di osservazioni

$$(1,4),\ (2,2),\ (3,6),\ (4,4),\ (5,6),\ (6,2),\ (7,4)\,.$$

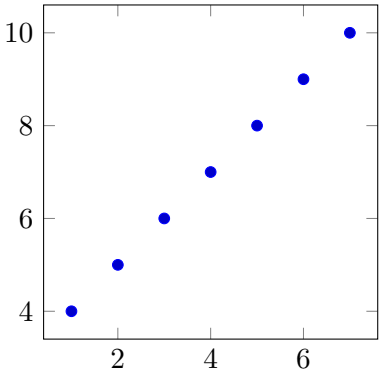
La covarianza è uguale a $s_{xy} = 0$, il coefficiente di correlazione è $r = 0$ ed il diagramma di dispersione è



Esempio 1.8.5. Su una popolazione di 7 unità, si considerino le coppie di osservazioni

$$(1,4),\ (2,5),\ (3,6),\ (4,7),\ (5,8),\ (6,9),\ (7,10)\,.$$

La covarianza è uguale a $s_{xy} = 4,66$, il coefficiente di correlazione è $r = 1$ ed il diagramma di dispersione è



Osservazione 1.8.6. Il software LibreOffice permette di calcolare molto facilmente covarianza e correlazione mediante i comandi COVAR e CORREL, entrambi di facile ed intuitivo utilizzo.

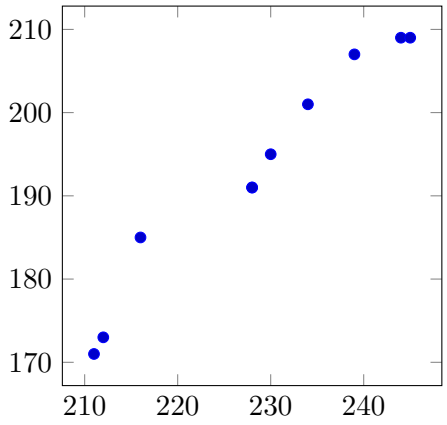
Esempio 1.8.7. I coefficienti di correlazione delle variabili numeriche riportate nella Tabella 1, calcolati con LibreOffice, sono:

	Pt	V	P	S	GF	GS	DR
Pt	1						
V	0,97	1					
P	-0,21	-0,42	1				
S	-0,94	-0,85	-0,13	1			
GF	0,84	0,86	-0,42	-0,70	1		
GS	-0,88	-0,83	0,01	0,90	-0,54	1	
DR	0,96	0,97	-0,64	-0,70	0,76	-0,77	1

Non è affatto sorprendente scoprire che la maggiore correlazione diretta si ha tra il numero di punti ed il numero di vittorie. Sono invece interessanti da un punto di vista statistico: l'alta correlazione diretta tra differenza reti e vittorie, la correlazione praticamente nulla tra pareggi e goal subiti, l'alta correlazione inversa tra punti e goal subiti.

Esempio 1.8.8. È molto ragionevole attendersi una forte correlazione diretta tra la progressioni temporali dei record mondiali di salto in alto maschile e femminile, essendo entrambi dipendenti in gran parte degli stessi fattori (tecniche di salto, metodologie di allenamento, controlli antidoping ecc.). Per quantificare tale correlazione elenchiamo in una tabella le misure, in centimetri, di entrambi i record agli inizi dei quinquenni dal 1950 al 1995 (dal 1 gennaio 1995 non vi sono nuovi record):

anno	maschile	femminile
1950	211	171
1955	212	173
1960	216	185
1965	228	191
1970	228	191
1975	230	195
1980	234	201
1985	239	207
1990	244	209
1995	245	209



Il coefficiente di correlazione è $r = 0,981$, quindi molto vicino a $+1$ ed in linea con le attese.

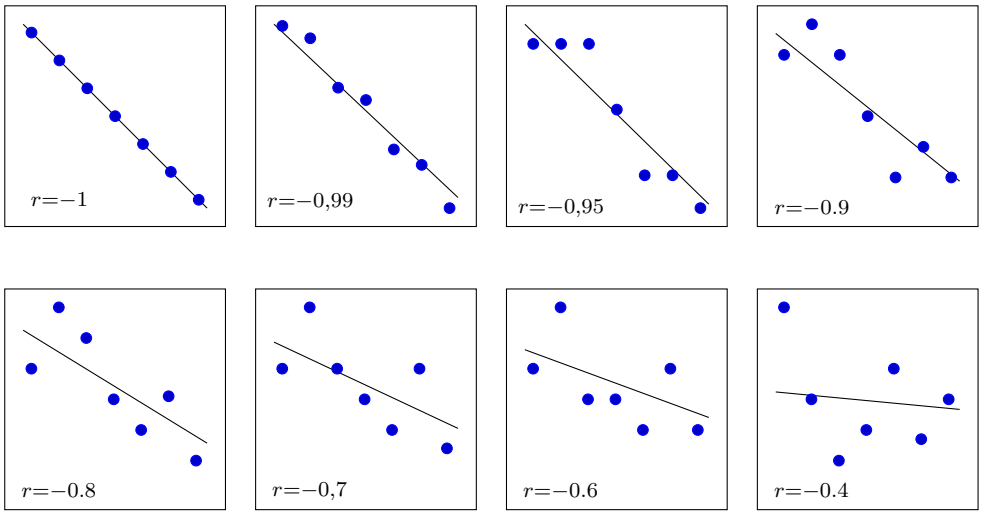
Come l'esempio precedente dimostra, l'esistenza di una correlazione tra due quantità misurate non implica una relazione diretta di causa-effetto. Lo studente deve avere piena coscienza di questo fatto per non cadere nelle trappole dialettiche dei soliti opinionisti da salotto che sfruttano l'analfabetismo scientifico a loro tornaconto. Spesso l'esistenza di una correlazione implica solamente che le variabili dipendono da una molteplicità di cause di cui alcune in comune, che a loro volta influenzano molte altre variabili. Inoltre, non sono infrequenti correlazioni tra fenomeni del tutto estranei tra loro.

Definizione 1.8.9. Date n osservazioni congiunte (x_1, y_1), (x_2, y_2),..., (x_n, y_n) di due variabili numeriche x e y, si dice **retta di regressione** la retta di equazione

$y = ax + b$, dove

$$a = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad b = \bar{y} - a\bar{x}\,.$$

Abbiamo già anticipato che, quando il coefficiente di correlazione è vicino a ± 1, il diagramma di dispersione tende ad addensarsi intorno alla retta di regressione, mentre quando r è vicino allo 0 la retta di regressione non risulta particolarmente significativa. Poiché la dimostrazione matematica di questo fatto può apparire alquanto ostica e misteriosa, è utile illustrare graficamente alcuni esempi numerici relativi a diversi valori di r. In ciascun esempio, al diagramma di dispersione (pallini blu) viene sovrapposta la retta di regressione.



Dal punto di vista matematico, la Definizione 1.8.9 può essere spiegata nel modo seguente. Se i punti (x_1, y_1), (x_2, y_2),..., (x_n, y_n) appartengono tutti alla retta di equazione $y = ax + b$, allora anche il loro baricentro $(\bar{x}, \bar{y})$ appartiene alla medesima retta. Infatti, sommando le n relazioni $y_i = ax_i + b$ e dividendo per n si ottiene

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\sum (ax_i + b)}{n} = a\bar{x} + b\,.$$

Poi, sottraendo la relazione $\bar{y} = a\bar{x} + b$ alle relazioni $y_i = ax_i + b$, otteniamo che i coefficienti a, b soddisfano le equazioni

$$b = \bar{y} - a\bar{x}, \qquad a(x_i - \bar{x}) - (y_i - \bar{y}) = 0, \qquad i = 1, \dots, n\,.$$

Se i punti (x_1, y_1), (x_2, y_2),..., (x_n, y_n) non sono allineati, allora le n condizioni $a(x_i - \bar{x}) - (y_i - \bar{y}) = 0$ non possono essere tutte verificate e si prende il valore di a che risulta essere il "meno peggio"; ciò avviene scegliendo a come il punto di minimo assoluto per la funzione

$$f(t) = \sum_{i=1}^n (t(x_i - \bar{x}) - (y_i - \bar{y}))^2\,.$$

Siccome

$$f'(t) = \sum_{i=1}^{n}(x_i - \bar{x})((t(x_i - \bar{x}) - (y_i - \bar{y})) = t\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

si ha $f'(t) = 0$ se e solo se

$$t = a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.$$

Il termine b viene poi preso in modo tale che il baricentro $(\bar{x}, \bar{y})$ sia contenuto nella retta di regressione.

Osserviamo anche (senza dimostrazione) che i coefficienti a, b della retta di regressione coincidono con quelli che minimizzano la quantità

$$\sum_{i=1}^{n}(ax_i + b - y_i)^2\,.$$

Esercizi.

Esercizio 1.16. Calcolare covarianza, correlazione e retta di regressione della successione di 7 misure bidimensionali:

$$(1,4),\ (2,1),\ (3,5),\ (4,3),\ (5,6),\ (6,2),\ (7,4)\,.$$

Esercizio 1.17. In relazione alla Tabella 1, senza eseguire calcoli e motivando la risposta, dire quanto vale il coefficiente di correlazione tra le variabili Pt e 2V-S.

Esercizio 1.18. È un fatto ben noto che se nelle principali città degli Stati Uniti d'America (dove le persone sono tradizionalmente più propense al trasferimento) si confronta la quantità di inquinanti nell'aria e la quantità di decessi per malattie respiratorie si scopre una correlazione inversa. Sapete spiegare questo fatto?

Esercizio 1.19. In una celebre trasmissione televisiva, un tizio afferma di essere guarito da una certa malattia bevendo ogni giorno una pozione a base di noccioli di pesca. La comunità scientifica, obbligata ad occuparsi del caso a scapito di altre più utili ricerche, deve stabilire quali delle seguenti ipotesi è statisticamente più vicina alla realtà:

(1) i noccioli di pesca hanno accelerato la guarigione;
(2) i noccioli di pesca non hanno avuto alcun effetto sulla guarigione;
(3) i noccioli di pesca hanno rallentato la guarigione.

Discutere come fare per risolvere il dilemma in maniera obiettiva e scientifica.

1.9. I test di ipotesi

Un altro campo di applicazione della statistica è quello di decidere in quale misura una certa ipotesi sulla popolazione, chiamata in gergo **ipotesi nulla**, è avvalorata dalle osservazioni su un campione statistico.

Ad esempio, un salumiere può ipotizzare che il peso medio delle fette di mortadella che escono dalla sua affettatrice sia di 25 grammi. Con un test su un campione di riferimento si può verificare se tale ipotesi è confermata o confutata dalla bilancia.

Ecco altri esempi di ipotesi nulle:

Popolazione	**Ipotesi nulla**
Lanci di una moneta	Testa esce con probabilità 50%
Residenti nel Lazio	Il 40% possiede il gruppo sanguigno 0
Abitanti di Londra nel 1636	Oltre metà dei decessi per peste
Funghi porcini	Tempo di digestione di almeno 4 ore

Oltre alla ipotesi nulla, conviene talvolta introdurre l'**ipotesi alternativa**, che contraddice totalmente l'ipotesi nulla: si ha dunque che solo una tra le ipotesi nulla ed alternativa è corretta. Solitamente l'ipotesi nulla viene indicata con la lettera H_0 e l'ipotesi alternativa con la lettera H_1.

È importante osservare che, in generale, i test su un campione non dimostrano la verità o la falsità di una ipotesi statistica, ma danno solo una indicazione probabilistica se tale ipotesi è avvalorata dalle osservazioni disponibili. Tale indicazione può essere più o meno significativa a seconda della consistenza del campione, della sua selezione, delle metodologie di misurazione eccetera.

Per determinare il supporto dato da un test all'ipotesi nulla viene introdotto un numero reale detto **livello di significatività**. *Maggiore è il livello di significatività di un test e maggiore è il supporto del test all'avvaloramento dell'ipotesi.*

La quantificazione del livello di significatività di un test avviene mediante l'analisi probabilistica degli errori; prima di procedere è utile soffermarci sul seguente esempio:

Esempio 1.9.1. L'allenatore dell'Atletico Sapienza vuol capire se il calciatore Gaudinho appena arrivato è abile a battere i rigori. A tal fine gli fa battere una serie di 20 rigori contro il portiere titolare e prende nota di quanti realizzati. Supponiamo, per fissare le idee, che Gaudinho abbia segnato 17 rigori su 20: niente male!

L'allenatore però non è ancora convinto: gli viene il dubbio che il portiere non è al massimo della condizione, che il pallone non è gonfiato al modo giusto, che il vento tira in maniera irregolare eccetera. Allora effettua un test sui migliori rigoristi (certificati) presenti in squadra: a ciascuno di loro fa battere una serie di 20 rigori e assume come livello di significatività la percentuale di coloro che ne segnano meno di 17.

Ogni test è soggetto a due tipi di errore: gli **errori del primo tipo** si hanno quando l'ipotesi nulla è vera ma viene rifiutata dal test, gli **errori del secondo tipo** si hanno quando l'ipotesi nulla è falsa ma viene accettata dal test.

Per esempio, supponiamo che l'ipotesi nulla H_0 sia la presenza del virus HIV nel sangue del signor J.; a seconda delle effettive condizioni di salute (prima colonna) e del risultato del test HIV (prima riga) si hanno le seguenti possibilità:

	test HIV positivo	**test HIV negativo**
J. sieropositivo (H_0 vera)	decisione corretta	errore del I tipo
J. sieronegativo (H_0 falsa)	errore del II tipo	decisione corretta

Gli errori del primo e del secondo tipo sono in generale molto diversi tra loro. Ad esempio, nel caso limite di un laboratorio di analisi del sangue in cui i test HIV danno sempre risultati positivi, non vengono mai commessi errori del I tipo e vengono commessi errori del II tipo in percentuale identica a quella dei pazienti non infettati dal virus.

Talvolta, per ridurre gli errori del I tipo è necessario aumentare quelli del II tipo e viceversa; in tal caso la decisione su come bilanciare le due tipologie viene fatta in base alla natura del problema con considerazioni di tipo etico, civile, economico, scientifico, sanitario ecc.

Esempio 1.9.2. Un tipico esempio di bilanciamento si ha nei test di autovalutazione delle conoscenze in ingresso.

Prima dell'iscrizione ad un corso di laurea della Facoltà di Scienze MM.FF.NN. gli studenti devono sottoporsi ad un test di domande di matematica di base: il test viene considerato superato se si risponde correttamente ad almeno N quesiti, con N numero fissato da apposita commissione (nel 2015 era $N = 11$).

La scelta di N viene fatta in modo da bilanciare nella maniera ritenuta ottimale gli errori di I e II tipo. Se l'ipotesi nulla da testare è la preparazione adeguata dello studente, allora alti valori di N aumentano la probabilità che uno studente preparato venga respinto al test (errore I tipo), mentre bassi valori di N aumentano la probabilità che uno studente non preparato venga promosso al test (errore II tipo).

Definizione 1.9.3. Il **livello di significatività** di un test è la probabilità che si ha nel commettere l'errore del I tipo, ossia di rifiutare un'ipotesi vera.

Il livello di significatività si esprime con un numero α compreso tra 0 e 1, oppure con il corrispondente valore percentuale: si può quindi dire equivalentemente che il livello di significatività è $0,05$ oppure del 5%. In pratica, bastano poche divergenze dovute al caso affinché il livello di significatività assuma valori prossimi allo 0 e di conseguenza ogni livello di significatività superiore a $0,05$ viene, con le dovute eccezioni, considerato come un buon supporto dei dati all'ipotesi nulla.

1.10. Il test del χ^2 (chi-quadro) di adattamento

In questa sezione ci occupiamo di un test di ipotesi ampiamente usato nella pratica scientifica per stabilire se un campione di dati osservati si adatta a una distribuzione teorica assegnata. Ad esempio, potrebbe esserci motivo di credere che il numero di incidenti sulle autostrade A1, A2 ecc. sia direttamente proporzionale alla loro lunghezza, oppure che il lancio di una moneta non perfettamente simmetrica e bilanciata dia testa nel 51% dei casi.

Supponiamo di avere un campione di n osservazioni di una variabile, raggruppate in una tabella di frequenze assolute contenente k classi, con $k \geq 2$ e n sufficientemente grande rispetto a k. Le classi possono rappresentare:

(1) caratteristiche qualitative (ad esempio il risultato di un'elezione);
(2) valori assunti da una variabile discreta (ad esempio il numero di incidenti);
(3) intervalli di valori assunti da una variabile continua (ad esempio il peso alla nascita).

Per ciascuna classe supponiamo di avere, oltre alla frequenza osservata f_i, $i = 1, \ldots, k$, una frequenza attesa a_i, con cui si vuole confrontare la frequenza osservata; le frequenze attese sono quelle che si osserverebbero se i dati del campione fossero distribuiti esattamente secondo la distribuzione ipotizzata.

Ad esempio, se l'ipotesi nulla riguarda "l'onesta" di un dado a sei facce, su n lanci le frequenze attese sono di $n/6$ per ciascuna faccia.

Se tutte le frequenze attese a_i sono sufficientemente alte (empiricamente almeno 5), per valutare quantitativamente la bontà dell'adattamento delle frequenze osservate alle frequenze attese si calcola la quantità

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - a_i)^2}{a_i},$$

che viene detta il **chi-quadro** calcolato dal campione. È evidente che tanto maggiore è il chi-quadro, tanto maggiore è la distanza del campione dalle frequenze attese.

Un basso valore di χ^2 vuol dire che il campione avvalora la validità dell'ipotesi attesa. Un alto valore di χ^2 vuol dire che il campione avvalora la non validità dell'ipotesi.

Per stabilire se il chi-quadro è basso oppure alto lo si confronta con un numero χ^2_α, detto **valore critico**, che dipende dal livello di significatività α che si vuol dare al test, e dal numero *df* di **gradi di libertà** della distribuzione teorica attesa.

Il senso ed il calcolo del numero di gradi di libertà è questione alquanto complicata ed una trattazione approfondita va al di là degli obiettivi di queste note. Quello che possiamo dire è che si ha sempre $df \leq k-1$ e vale $df = k-1$ *quando il modello teorico è fissato prima di aver raccolto le frequenze assolute.*

La regola pratica è molto semplice: *se $\chi^2 \leq \chi^2_\alpha$ allora il campione avvalora la distribuzione teorica con livello di significatività α*. Detto in altri termini, se ci mettiamo d'accordo nel rifiutare un'ipotesi ogni volta che $\chi^2 > \chi^2_\alpha$, allora la probabilità di scartare un'ipotesi corretta è uguale ad α.

Anche il calcolo del valore critico, come funzione di α e *df*, richiede strumenti probabilistici qui non trattati, tuttavia tale valore è calcolato da molti software (compreso il già citato LibreOffice): una serie di valori critici sono riportati nella Tabella 2. Non sorprende affatto osservare che, a parità di gradi di libertà, il valore critico aumenta al diminuire di α.

Esempio 1.10.1. Per stabilire se nel lancio di una moneta si ha eguali probabilità di avere testa o croce, si effettua un test di 100 lanci. Le frequenze attese sono 50 e 50, i gradi di libertà sono $df = k - 1 = 2 - 1 = 1$ e se vogliamo avere un livello di significativià del 10% il valore critico da considerare è

$$\chi^2_{0,1} = 2,706\,.$$

Supponiamo che nei 100 lanci si ottenga 55 volte testa e 45 volte croce: il chi-quadro del campione è allora

$$\chi^2 = \frac{(55-50)^2}{50} + \frac{(45-50)^2}{50} = \frac{25}{50} + \frac{25}{50} = 1\,.$$

Siccome

$$1 = \chi^2 \leq \chi^2_{0,1} = 2,706$$

il test avvalora l'ipotesi. Allo stesso modo, siccome $\chi^2_{0,9} = 0,016$ il medesimo test non avrebbe avvalorato l'ipotesi nulla se avessimo richiesto un livello di significatività del 90%.

Esempio 1.10.2. Eseguiamo 1000 lanci di una moneta. Vogliamo individuare quante volte deve uscire testa per avere avvalorata, con un livello di significativià del 90%, l'ipotesi che testa e croce siano equiprobabili. Come nell'esempio precedente si ha un solo grado di libertà ed il valore critico, ricavato dalla Tabella 2, è uguale

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

TABELLA 2. Valori critici per i livelli di significatività maggiormente usati in pratica: il puntino separa la parte intera dalla parte decimale.

a $\chi^2_{0,9} = 0,016$. Se indichiamo con x la differenza tra il numero di volte in cui esce testa ed il valore atteso di 1000 si ha

$$\chi^2 = \frac{x^2}{500} + \frac{x^2}{500} = \frac{x^2}{250}$$

e quindi l'ipotesi di equiprobabilità è avvalorata se

$$\frac{x^2}{250} \leq 0,016$$

ossia se $x^2 \leq 250 \cdot 0,016 = 4$, o equivalentemente se esce testa un numero di volte compreso tra 498 e 502.

1.11. Il test del χ^2 (chi-quadro) di indipendenza

Il test del chi-quadro si usa non solo per avvalorare un'ipotesi teorica ma anche per stabilire l'indipendenza di due variabili osservate su una medesima popolazione. Come vedremo, la procedura è molto simile al test di adattamento, con la differenza

sostanziale che se la prima variabile assume k valori e la seconda variabile h valori, allora il numero di gradi di libertà è uguale a

$$df = (k-1)(h-1)\,.$$

Ad esempio, per smontare le tesi di qualche buontempone che ha studiato alla Iutubb Universtity, si vuol verificare che tra i laureati in Scienze Naturali, il voto di Laurea è indipendente dal fattore Rh. Per semplicità espositiva dividiamo i possibili voti di laurea in tre classi, la prima con i voti da 105 a 110 e lode, la seconda con i voti tra 95 e 104, la terza con i voti tra 66 e 94.

Scelto un campione abbanstanza significativo di n laureati, si avrà una tabella delle frequenze sotto forma di matrice:

	Rh positivo	Rh negativo
105-110L	o_{11}	o_{12}
95-104	o_{21}	o_{22}
66-94	o_{31}	o_{32}

dove o_{11} è il numero dei laureati con Rh positivo e voto di Laurea compreso tra 105 e 110 e lode, o_{32} il numero dei laureati con voto basso ed Rh negativo eccetera. Se indichiamo con o_{*1} il numero di laureati con Rh positivo e con o_{*2} quello dei laureati con Rh negativo, si ha

$$o_{*1} = o_{11} + o_{21} + o_{31}, \quad o_{*2} = o_{12} + o_{22} + o_{32}, \qquad n = o_{*1} + o_{*2}\,.$$

Il numero dei laureati con voti alti, medi e bassi è rispettivamente uguale a:

$$o_{1*} = o_{11} + o_{12}, \qquad o_{2*} = o_{21} + o_{22}, \qquad o_{3*} = o_{31} + o_{32}\,.$$

L'ipotesi che le due caratteristiche siano indipendenti implica che, *a meno di fluttuazioni statistiche*, la proporzione tra o_{11} ed o_{12} è la stessa tra o_{*1} ed o_{*2}: si ha quindi un valore atteso per gli studenti con voto alto e RH positivo di

$$a_{11} = o_{1*} \cdot \frac{o_{*1}}{n} = \frac{o_{1*} \cdot o_{*1}}{n}$$

e similmente per ogni i, j

$$a_{ij} = \frac{o_{i*} o_{*j}}{n} = \frac{1}{n} \Big(\sum \text{frequenze riga } i \Big) \Big(\sum \text{frequenze colonna } j \Big)\,.$$

Adesso, se tutte le frequenze attese a_{ij} sono sufficientemente alte (almeno 5) si procede come nel test di adattamento, ossia si calcola il chi-quadro del campione

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - a_{ij})^2}{a_{ij}}$$

e si guarda se tale valore supera o meno il valore critico.

Per capire meglio, facciamo un esempio numerico (con dati di fantasia) su un campione ipotetico di 220 laureati e con la matrice 3×2 delle frequenze osservate uguale a

o_{ij}	Rh positivo	Rh negativo	totale
105-110L	65	32	97
95-104	51	26	77
66-94	31	15	46
totale	147	73	220

La corrispondente matrice delle frequenze attese è:

a_{ij}	Rh positivo	Rh negativo
105-110L	$\frac{97\cdot 147}{220} = 64,8$	$\frac{73\cdot 97}{220} = 32,2$
95-104	$\frac{77\cdot 147}{220} = 51,4$	$\frac{73\cdot 77}{220} = 25,6$
66-94	$\frac{46\cdot 147}{220} = 30,7$	$\frac{73\cdot 46}{220} = 15,3$

che fornisce un valore del chi-quadro uguale a

$$\chi^2 = \frac{(0,2)^2}{64,8} + \frac{(-0,2)^2}{32,2} + \frac{(-0,4)^2}{51,4} + \frac{(0,4)^2}{25,6} + \frac{(0,3)^2}{30,7} + \frac{(-0,3)^2}{15,3} = 0,02\,.$$

La matrice è 3×2 e quindi i gradi di libertà sono $(3-1)(2-1) = 2$. Il valore critico su due gradi di libertà e livello di significatività del 99% è uguale a $0,02$. Possiamo quindi affermare che sul campione analizzato le due variabili sono indipendenti ad un livello di significatività del 99%.

Riepilogando, se su un campione di n unità si osservano due variabili, la prima che assume k valori e la seconda che ne assume h, per valutare quanto tali variabili sono indipendenti a meno di fluttuazioni statistiche, si procede ne modo seguente:

(1) le frequenze assolute osservate o_{ij} vengono disposte in una matrice $k \times h$;
(2) si calcola la matrice delle frequenze attese a_{ij} in ipotesi di indipendenza. Il coefficiente a_{ij} è uguale a
$$a_{ij} = \frac{1}{n}(o_{i1} + o_{i2} + \cdots + o_{ih})(o_{1j} + o_{2j} + \cdots + o_{kj})\,;$$
(3) se ogni a_{ij} è abbastanza alto (empiricamente almeno 5) si calcola il chi-quadro
$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - a_{ij})^2}{a_{ij}}\,;$$
(4) si confronta il chi-quadro con il valore critico χ^2_α a $(k-1)(h-1)$ gradi di libertà. Se $\chi^2 \leq \chi^2_\alpha$ il test di indipendenza ci autorizza a dire che le due variabili sono indipendenti sul campione osservato ad un livello di significatività superiore ad α.