Cenni di statistica descrittiva

- La **statistica descrittiva** è la disciplina nella quale si studiano le metodologie di cui si serve uno sperimentatore per *raccogliere*, *rappresentare* ed *elaborare* dei dati osservati ai fini dell'analisi di un certo fenomeno.

- Tale disciplina è distinta dalla **statistica inferenziale** *(o induttiva)* che studia le metodologie che permettono di generalizzare ed estendere alla popolazione le informazioni ottenute da un indagine campionaria.

- In particolare nella statistica inferenziale giocano un ruolo determinante le metodologie probabilistiche.

Popolazione. E' l'insieme i cui elementi, dette **unità statistiche**, hanno in comune almeno una caratteristica.

Tali caratteristiche possono essere di tipo *qualitativo*, oppure *numerico*.

Tabella 1: Esempi di popolazione

Popolazione	unità statistica	caratteristica
Nati a Roma nel 2003	bambino	sesso (Qual.)
Studenti di Ingegneria	persona	altezza, età (Num.)
Giorni dell'anno	giorno	temperatura (Num.)

In generale si parla di **caratteri** o **attributi** (presenti, eventualmente in un certo grado, o assenti negli elementi della popolazione);

Noi ci occuperemo solo di caratteristiche di tipo numerico per le quali si usa il il termine di **variabile**.

Pertanto le popolazioni oggetto di studio sono costituite da un insieme di numeri che costituiscono la misurazione della caratteristica comune agli elementi della popolazione in oggetto.

La statistica descrittiva si articola in 3 fasi fondamentali. La rilevazione, la rappresentazione e l'elaborazione dei dati.

La rilevazione dei dati. Acquisire le informazioni sul fenomeno collettivo. Schematicamente consiste in:

- Descrizione del fenomeno oggetto dell'indagine.
- Individuazione della popolazione e delle unità statistiche che la compongono.
- Determinazione dei caratteri (aspetti del fenomeno da rilevare).
- Raccolta dei dati.
- Spoglio (conteggio, ordinamento e classificazione).

La rappresentazione dei dati. Rappresentare mediante grafici o tabelle le caratteristiche dei dati rilevati.

L'elaborazione dei dati. Ottenere degli indici di sintesi sui dati rilevati e studiare relazioni statistiche tra gli stessi.

Ordinamento e frequenze

I dati grezzi raccolti nella fase di rilevazione, ad esempio

$$z_1 \ , \quad z_2 \ , \quad \cdots \ , \quad z_r \ ,$$

sono generalmente di difficile interpretazione, per cui una prima operazione utile consiste nell'elencare i dati stessi secondo grandezza, ad esempio in ordine crescente :

$$y_1 \ \leq \ y_2 \ \leq \ \cdots \ \leq \ y_r \ .$$

In questo modo possiamo determinare il *rango* o *campo di variazione* dei dati, rappresentato dalla differenza tra il più grande e il più piccolo, cioè

$$rango \ = \ y_r - y_1 \, .$$

I dati numerici raccolti potranno essere in parte (o anche tutti) coincidenti e quindi, indicando con

$$x_1 \, , \, x_2 \, , \, \ldots \, , \, x_n \, ,$$

i valori distinti, si ha $y_j \in \{x_1, \, x_2, \, \ldots \, , \, x_n\}$ per ogni $j = 1, \ldots , r$, con $n \leq r$.

Se, per ogni $i = 1, 2, \ldots , n$, indichiamo con r_i il numero di dati uguali a x_i, si ha

$$r_1 + r_2 + \cdots + r_n = r \ .$$

I valori $r_1, \ldots , r_n$ sono le *frequenze assolute* con cui si presentano i dati $x_1, \ldots , x_n$, mentre i valori

$$f_1 = \frac{r_1}{r} \ , \quad f_2 = \frac{r_2}{r} \ , \quad \ldots \ , \quad f_n = \frac{r_n}{r}$$

sono le *frequenze relative*.

Ovviamente : $f_1 + f_2 + \cdots + f_n = 1$.

Se i dati sono ordinati si può definire la *frequenza cumulata*, riferita alla modalità x_k, come la somma delle frequenze (assolute per la frequenza cumulata assoluta, relative per la frequenza cumulata relativa) dalla prima modalità x_1 fino a x_k.
Cioè, la frequenza cumulata assoluta di x_k è data da

$$R_k = r_1 + \cdots + r_k \ , \quad k = 1, 2, \ldots, n \ ,$$

e la frequenza cumulata relativa di x_k è data da

$$F_k = f_1 + \cdots + f_k \ , \quad k = 1, 2, \ldots, n \ .$$

Esempio 1 Supponiamo che in una classe di 28 ragazzi sia stato proposto il quesito: *quale sport preferisci?* Se le risposte sono state: il calcio (10 ragazzi), il tennis (4 ragazzi), la pallacanestro (6 ragazzi), il nuoto (3 ragazzi), altro (5 ragazzi), possiamo riassumere l'indagine con la tabella delle frequenze assolute e delle frequenze relative (Tabella 2), riportata sotto.

Tabella 2: Frequenze ass. e relative

sport	freq. ass.	freq. rel.
calcio	10	0.36
tennis	4	0.14
pallacanestro	6	0.21
nuoto	3	0.11
altro	5	0.18
totale	28	1

In questo esempio non potendo ordinare i dati non si possono calcolare le frequenze cumulate.

Esempio 2 Consideriamo la popolazione dell'esercizio precedente, alla quale stavolta viene formulata la domanda *Quanti anni hai?*. Se le risposte sono state: 18 (4 persone), 19 (13 persone), 20 (7 persone), 21 (3 persone), 22 (1 persona), si ottiene la seguente tabella.

Tabella 3: Frequenze semplici e cumulate

età	fr.ass.	fr. cum. ass.	fr. rel.	fr. cum. rel.
18	4	4	0.14	0.14
19	13	17	0.46	0.61
20	7	24	0.25	0.86
21	3	27	0.11	0.96
22	1	28	0.04	1
totale	28		1	

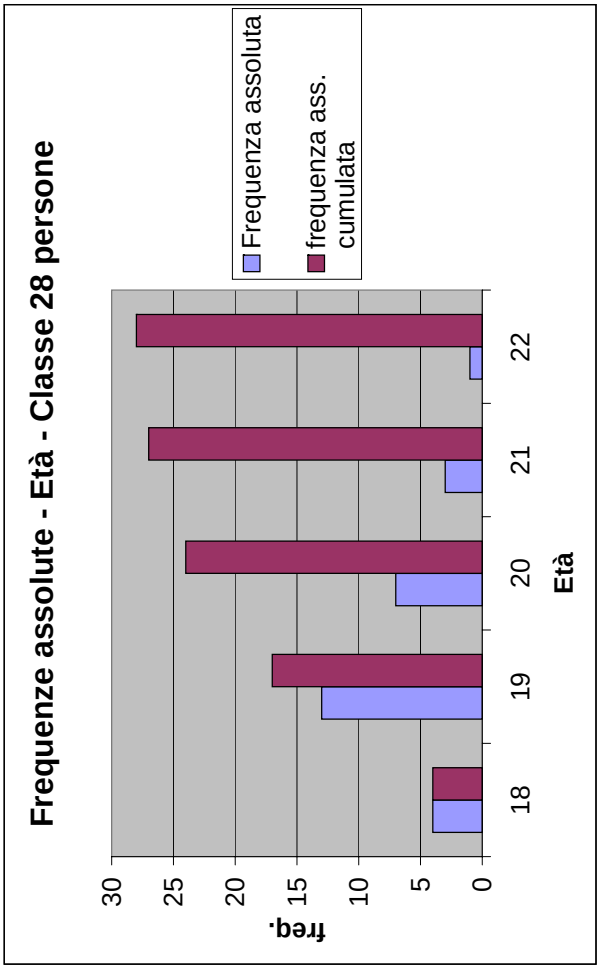


Figura 1: Frequenza assoluta semplice e cumulata

Se l'insieme di dati da studiare è troppo grande si può pensare di raggrupparli in *classi*.

Ad esempio, considerando una variabile X che assume valori in un intervallo $[a, b]$, una suddivisione in classi consiste nel dividere $[a, b]$ in intervalli disgiunti (in genere di uguale ampiezza)

$$[a_0, a_1) \ , \quad [a_1, a_2) \ , \quad \dots \ , \quad [a_{m-1}, a_m] \ ,$$

con

$$a_0 = a \ < \ a_1 \ < \ \cdots \ < \ a_m = b \ .$$

I dati vengono raggruppati nelle rispettive classi di appartenenza, calcolando le *frequenze di classe assolute* $n_1, \dots, n_m$, *oppure le* frequenze di classe relative $p_1, \dots, p_m$.

La frequenza n_k rappresenta il numero di dati appartenenti all'intervallo $[a_{k-1}, a_k)$, mentre la frequenza relativa p_k è pari al rapporto $\frac{n_k}{r}$, dove r è il numero dei dati osservati.

Il numero delle classi deve essere scelto in modo che non siano nè troppe (nel qual caso in ogni classe ci sarebbero pochissimi dati) nè troppo poche (nel qual caso si avrebbero molti elementi in poche classi e la

rappresentazione risultante non sarebbe significativa in quanto avremmo perso troppa informazione sulla distribuzione reale).

In genere si sceglie un numero (intero) di classi prossimo al valore $1 + \frac{10}{3} Log_{10} r$*. I valori delle frequenze (assolute, relative, cumulate, cumulate relative) possono poi essere riportati in corrispondenti* tabelle di frequenza*.*

Esempio 3 I risultati ottenuti da 74 studenti durante un test (il voto massimo è di 250) sono riportati nella Tabella 4

Tabella 4: Voti dei 74 studenti

65	158	114	183	124	94
76	203	120	145	177	123
81	121	150	90	137	213
25	186	103	105	194	129
36	40	164	55	173	213
103	97	246	200	159	67
144	106	238	218	156	147
73	108	46	230	151	148
184	89	111	206	157	126
64	118	151	236	137	237
84	196	134	205	187	148
149	185	132	160	168	143
155	161				

Raggruppiamo i dati in classi e costruiamo una tavola che riporti le frequenze, le frequenze cumulative e quelle relative

Secondo la regola suggerita si ricava un numero di classi arrotondato per eccesso pari a 8

($1 + \frac{10}{3} Log_{10} 74 = 7.23$), ciascuna di ampiezza 28, ottenendo la suddivisione riportata nella Tabella 5.

Tabella 5: Frequenze Esempio 3

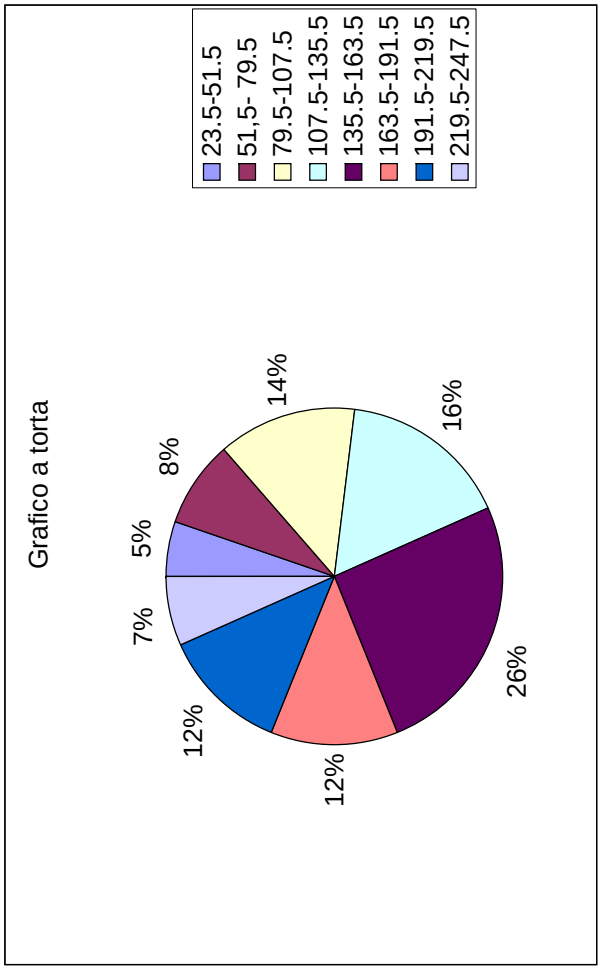
Classi	Centro di classe	Fr. di classe	Fr. rel.	Fr. cum.	Fr. cum. rel.
23.5- 51.5	37.5	4	0.054	4	0.054
51,5- 79.5	65.5	6	0.081	10	0.135
79.5-107.5	93.5	10	0.135	20	0.270
107.5-135.5	121.5	12	0.162	32	0.432
135.5-163.5	149.5	19	0.257	51	0.689
163.5-191.5	177.5	9	0.122	60	0.811
191.5-219.5	205.5	9	0.122	69	0.932
219.5-247.5	233.5	5	0.067	74	1.000

Le tabelle di frequenza pur contenendo molte informazioni non consentono di cogliere a colpo d'occhio eventuali peculiarità presenti nei dati.

Ciò è invece reso possibile dai diversi metodi di rappresentazione grafica*.*

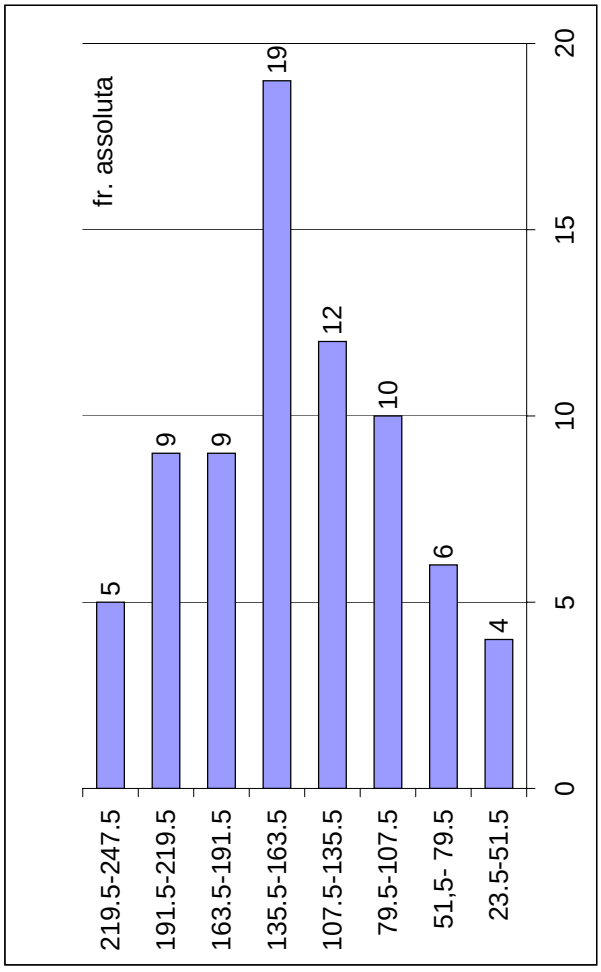
Di seguito ne elenchiamo alcuni fra quelli più usati.

- diagrammi a torte*: si divide un cerchio in settori circolari che rappresentano le categorie considerate. Ogni settore ha un'ampiezza proporzionale alla frequenza della corrispondente categoria.*



- grafi a barre*: ogni raggruppamento è rappresentato da una barra la cui lunghezza è proporzionale alla corrispondente frequenza.*

  Tali diagrammi sono usati di solito per i fenomeni di tipo qualitativo, nei quali non si possono effettuare misurazioni.



- istogrammi*: servono per rappresentare dati raggruppati in classi.*

  Si divide l'asse delle ascisse in intervalli contigui

di ampiezza uguale a quella delle corrispondenti classi e su ogni intervallo si riporta un rettangolino di area uguale alla frequenza della classe relativa.

Se si usano le frequenze assolute si parla di istogramma delle frequenze assolute e l'area totale dei rettangolini è pari al numero totale di osservazioni.

Nel caso delle frequenze relative si parla di istogramma delle frequenze relative e l'area totale dei rettangolini è pari a 1.

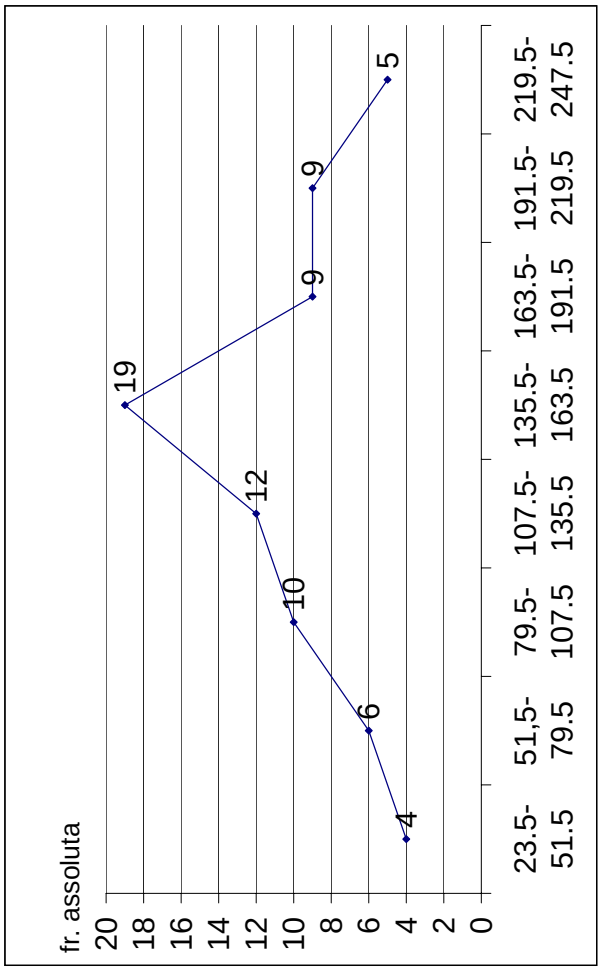
- poligoni di frequenza*:*

  si rappresentano i dati mediante una spezzata (anzichè un diagramma a scalini come avviene per gli istogrammi).

  Ogni classe è rappresentata dal suo valore centrale, riportando in corrispondenza un punto di ordinata uguale alla frequenza della classe.

Tali punti vengono poi uniti mediante segmenti.

In modo analogo, si possono costruire gli istogrammi e i poligoni di frequenza cumulata (assoluta oppure relativa), nel qual caso il diagramma è a forma di scalinata oppure di spezzata, entrambe monotone.



Misure descrittive

Allo scopo di presentare in forma chiara e sintetica le principali informazioni presenti nei dati occorre riassumere mediante opportune misure o indici numerici le rilevazioni effettuate.

Le misure impiegate più di frequente riguardano principalmente due aspetti:

- misure di posizione (o di tendenza centrale)*;*
- misure di dispersione (o di variazione)*.*

Misure di tendenza centrale

- Media aritmetica*: la media aritmetica* $\overline{z}$ *di un insieme di dati*

$$z_1 \ , \quad z_2 \ , \quad \dots \ , \quad z_r$$

è il numero

$$\overline{z} = \frac{1}{r} \sum_{i=1}^{r} z_i \ .$$

Utilizzando i valori distinti $x_1, \dots, x_n$ *e le rispettive frequenze relative* $f_1, \dots, f_n$*, la media aritmetica* $\overline{z}$ *si può anche esprimere come* media ponderata *nel seguente modo:*

$$\overline{z} = \sum_{k=1}^{n} f_k x_k \ .$$

1. *la media aritmetica $\overline{w}$ dell'insieme di dati*

$$w_1 = z_1 + a \ , \quad w_2 = z_2 + a \ , \quad \dots \ , \quad w_r = z_r + a \ ,$$

dove a è una costante reale, è il numero

$$\overline{w} = \overline{z} + a \ .$$

In particolare, se $a = -\overline{z}$ segue $\overline{w} = 0$.

2. *la media aritmetica $\overline{v}$ dell'insieme di dati*

$$v_1 = bz_1 \ , \quad v_2 = bz_2 \ , \quad \dots \ , \quad v_r = bz_r \ ,$$

dove b è una costante reale, è il numero

$$\overline{v} = b\overline{z} \ .$$

3. *la media aritmetica $\overline{u}$ dell'insieme di dati*

$$u_1 = bz_1 + a \ , \quad u_2 = bz_2 + a \ , \quad \dots \ , \quad u_r = bz_r + a \ ,$$

dove a e b sono due costanti reali, è il numero

$$\overline{u} = b\overline{z} + a \ .$$

4. *considerati due insiemi di dati (aventi uguale numerosità)*

$$z_1 \ , \ \ z_2 \ , \ \ \dots \ , \ \ z_r \ ,$$

$$w_1 \ , \ \ w_2 \ , \ \ \dots \ , \ \ w_r \ ,$$

e due costanti reali a, b, la media aritmetica $\overline{u}$ dell'insieme di dati

$$u_1 = az_1 + bw_1, \ u_2 = az_2 + bw_2, \ \dots, \ u_r = az_r + bw_r$$

è il numero

$$\overline{u} = a\overline{z} + b\overline{w} \ .$$

- Media geometrica*: la media geometrica $\overline{z}_g$ dei numeri*

$$z_1 \ , \ \ z_2 \ , \ \ \dots \ , \ \ z_r$$

è data da

$$\overline{z}_g = (z_1 z_2 \cdots z_r)^{\frac{1}{r}} \ .$$

Tale media risulta appropriata in situazioni simili a quelle descritte nei seguenti due problemi.

Esempio 4 In un periodo di 8 anni il tasso di interesse composto applicato sui depositi da una banca è stato : 7.1% per 2 anni, 7.9% per 3 anni, 7.5% per 2 anni e 7.4% per 1 anno. Qual'è il tasso medio annuo?

Indicando con $r_1, \ldots, r_8$ i tassi applicati negli 8 anni e con r il tasso medio, dev'essere

$$(1+r)^8 = (1+r_1)(1+r_2)\cdots(1+r_8),$$

e quindi $1+r$ è la media geometrica dei numeri

$$1+r_1\ ,\ \ \ldots\ ,\ \ 1+r_8\ ,$$

cioè

$$(1+r) = [(1+r_1)(1+r_2)\cdots(1+r_8)]^{\frac{1}{8}} \ = \ 1.0754\ .$$

Pertanto $r = 0.0754$, cioè il tasso medio applicato dalla banca è stato del 7.54%.

Esempio 5 Dato un parallelepipedo i cui lati misurano rispettivamente $8\,cm, 5\,cm$ e $25\,cm$, calcolare la lunghezza l del lato del cubo avente lo stesso volume.

Dev'essere ovviamente $l^3 = 8 \times 5 \times 25 = 1000$ e quindi $l = 10\,cm$, cioè l è la media geometrica delle misure dei lati del parallelepipedo.

- Media armonica*: la media armonica $\overline{z}_a$ dei numeri*

$$z_1\ ,\quad z_2\ ,\quad \cdots\ ,\quad z_r$$

è data da

$$\overline{z}_a = \frac{r}{\frac{1}{z_1} + \cdots + \frac{1}{z_r}}\ .$$

Di seguito esaminiamo un'applicazione della media armonica.

Esempio 6 Un'automobile ha percorso un tratto di strada alla velocità costante di $80Km/h$ all'andata e di $120Km/h$ al ritorno.

Qual'è, ai fini del tempo totale di percorrenza, la velocità media sull'intero percorso?

Come nell'esercizio precedente la media aritmetica darebbe un risultato errato, mentre la media che ci dà il valore esatto è in questo caso quella armonica.

Infatti, indicando con s la lunghezza del tratto di strada, i tempi di percorrenza (misurati in ore) all'andata e al ritorno sono rispettivamente

$$t_a = \frac{s}{80} \; ; \quad t_r = \frac{s}{120} \; .$$

Quindi il tempo totale t è dato da $t_a + t_r$ e la velocità media sull'intero percorso è (la media armonica delle due velocità date)

$$\overline{v}_a = \frac{2s}{t_a + t_r} = \frac{2s}{\frac{s}{80} + \frac{s}{120}} = \frac{2}{\frac{1}{80} + \frac{1}{120}} = 96 Km/h \; .$$

In questo esempio la media aritmetica $\overline{v}$ e la media geometrica $\overline{v}_g$ delle due velocità date risultano

rispettivamente 100 e $40\sqrt{6} \simeq 97.98$.

Come si può notare (e come si potrebbe dimostrare in generale, nel caso di dati numerici tutti positivi) risulta

media armon. $<$ *media geom.* $<$ *media aritm.* .

Le medie precedenti si possono inquadrare da un unico punto di vista adottando la seguente definizione (introdotta da Chisini in un lavoro del 1929 e ulteriormente generalizzata in un lavoro di de Finetti del 1931):

se di r *grandezze omogenee* $z_1, \ldots, z_r$ *interessa valutare una funzione simmetrica* $f(z_1, \ldots, z_r)$ *e per un certo valore* z^* *risulta*

$$f(z^*, \ldots, z^*) = f(z_1, \ldots, z_r) \ ,$$

il valore z^* *si dice media di* $z_1, \ldots, z_r$ *ai fini del calcolo di* f.

Infatti, per calcolare il valore di f tutto va come se fosse

$$z_1 \;=\; z_2 \;=\; \cdots \;=\; z_r \;=\; z^* \,.$$

Nel caso della media aritmetica la funzione f è la somma, *nel caso della media geometrica f è il* prodotto, *mentre nel caso della media armonica f è la* somma dei valori inversi*.*

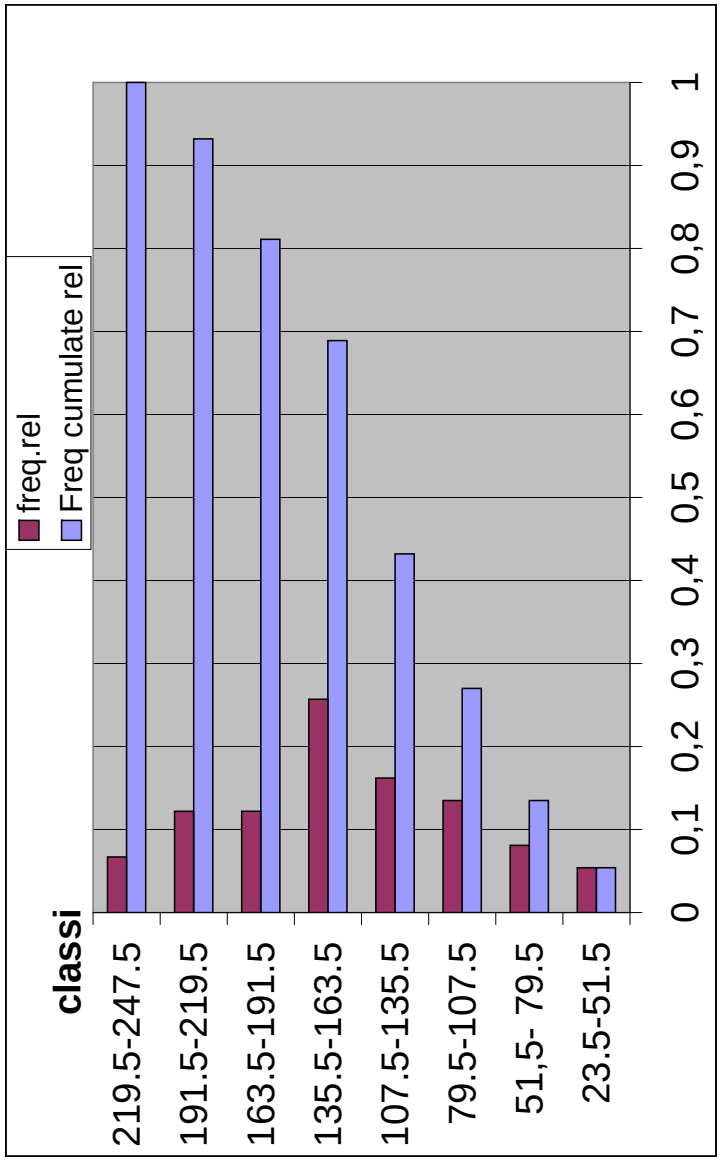
- Mediana*: la mediana di un insieme di numeri, ordinati in ordine crescente oppure decrescente,*

$$z_1 \;,\;\; z_2 \;,\;\; \cdots \;,\;\; z_r$$

è il valore centrale se r è dispari, altrimenti è la media aritmetica dei due valori centrali se r è pari.

La media aritmetica è fortemente influenzata dai valori estremi (in particolare dalla presenza di valori anomali), mentre la mediana non ne risente.

Pertanto, la mediana è preferibile nei casi in cui ci sono pochi dati sperimentali, oppure la gran parte dei dati sono concentrati verso un estremo.



- frattili: *si definisce frattile di ordine p quel valore*

alla sinistra del quale sta una frazione p dei dati.

Ad esempio la mediana è il frattile di ordine 50%.

Inoltre, si possono definire tre quartili Q_1, Q_2, Q_3, *come i valori che dividono l'insieme ordinato dei dati in quattro parti uguali.*

Alla sinistra di Q_1 stanno il 25% dei dati, alla sinistra di Q_2 (che coincide con la mediana) stanno il 50% dei dati, mentre alla sinistra di Q_3 stanno il 75% dei dati.

In modo analogo si possono definire i decili *e i* percentili.

- moda*: considerato l'insieme di dati*

$$z_1 \ , \ \ z_2 \ , \ \ \dots \ , \ \ z_r \ ,$$

siano

$$x_1 \ , \ \ x_2 \ , \ \ \dots \ , \ \ x_n$$

i valori distinti e

$$r_1 \ , \ \ r_2 \ , \ \ \ldots \ , \ \ r_n$$

le rispettive frequenze assolute.

Si definisce moda dell'insieme dei dati ogni valore che compare con frequenza massima, cioè ogni valore x_k *tale che* $r_k \geq r_i \ , \ \ i = 1, 2, \ldots, n$.

Quando i dati sono raggruppati in classi si possono individuare una o più classi modali, che corrispondono nell'istogramma ad altrettanti massimi.

La moda può risultare utile quando i dati sono divisi in classi che non sono di tipo numerico (ad esempio, luogo di nascita, professione, ...).
D'altra parte se la moda non è unica, la sua utilità appare limitata.

Osserviamo che per le distribuzioni di dati unimodali e simmetriche, com'è facile verificare, la media aritmetica, la mediana e la moda coincidono.

Misure di dispersione

Le misure di tendenza centrale non ci dicono nulla su come i dati sono distribuiti intorno al valore centrale.

Infatti due o più insiemi di dati possono avere uno stesso valore centrale e allo stesso tempo essere distribuiti in modo completamente differente intorno ad esso.

Per misurare la dispersione dei dati si introducono degli indici di variabilità.

In questo senso il rango o campo di variazione definito in precedenza è un primo indice di dispersione che, però, diventa poco significativo se uno dei dati è anomalo (cioè molto grande o molto piccolo).

Osserviamo anche che la media aritmetica $\overline{w}$ *delle deviazioni dalla media*

$$w_1 = z_1 - \overline{z}\ ,\ \ w_2 = z_2 - \overline{z}\ ,\ \ \ldots\ ,\ \ w_r = z_r - \overline{z}\ ,$$

non è utile per misurare la dispersione dei dati in quanto, come si è visto in precedenza, risulta sempre $\overline{w} = 0$.

Una misura che potrebbe essere utilizzata è la media dei valori assoluti *delle deviazioni dalla media, detta* deviazione media *e pari alla quantità*

$$\frac{1}{r}(|z_1 - \overline{z}| + |z_2 - \overline{z}| + \cdots + |z_r - \overline{z}|) \ .$$

La deviazione media non è facilmente trattabile dal punto di vista matematico, mentre risulta più conveniente considerare la media dei quadrati *delle deviazioni dalla media, che si dice* varianza *ed è definita dalla quantità*

$$\sigma_Z^2 = \frac{1}{r}\left[(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \cdots + (z_r - \overline{z})^2\right] \ .$$

Nel caso particolare $z_1 = z_2 = \cdots = z_r = z$ *risulta* $\overline{z} = z$ *e quindi* $\sigma_Z^2 = 0$. *Al contrario, se per almeno due indici* i, j *si ha* $z_i \neq z_j$, *allora* $\sigma_Z^2 > 0$.

La radice quadrata σ_Z della varianza si chiama scarto quadratico medio *o* deviazione standard *e rappresenta anch'essa una misura di dispersione dei dati. A differenza della varianza, però, la deviazione standard è espressa nelle stesse unità di misura dei dati.*

Proprietà della varianza

- *sviluppando i quadrati, la varianza si può rappresentare come differenza fra la* $media\ dei\ quadrati\ e\ il\ quadrato\ della\ media$

$$\sigma_Z^2 = \frac{1}{r}\left[z_1^2 + \cdots + z_r^2 - 2\overline{z}(z_1 + \cdots + z_r) + r\overline{z}^2\right] =$$

$$= \frac{1}{r}\sum_{i=1}^{r} z_i^2 - \left(\frac{1}{r}\sum_{i=1}^{r} z_i\right)^2 = \overline{z^2} - \overline{z}^2.$$

- *la varianza σ_W^2 dell'insieme di dati*

$$w_1 = z_1 + a\ ,\ \ w_2 = z_2 + a\ ,\ \ \ldots\ ,\ \ w_r = z_r + a$$

coincide con σ_Z^2.

Infatti, essendo $w_i - \overline{w} = z_i - \overline{z}$, *si ha*

$$\sigma_W^2 = \tfrac{1}{r}\left[(w_1 - \overline{w})^2 + (w_2 - \overline{w})^2 + \cdots + (w_r - \overline{w})^2\right] =$$

$$= \tfrac{1}{r}\left[(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \cdots + (z_r - \overline{z})^2\right] = \sigma_Z^2 \ .$$

- *la varianza* σ_U^2 *dell'insieme di dati*

$$u_1 = bz_1 \ , \ \ u_2 = bz_2 \ , \ \ \ldots \ , \ \ u_r = bz_r$$

è uguale a $b^2\sigma_Z^2$.
Infatti, essendo $u_i - \overline{u} = b(z_i - \overline{z})$, *si ha*

$$\sigma_U^2 = \tfrac{1}{r}\left[(u_1 - \overline{u})^2 + (u_2 - \overline{u})^2 + \cdots + (u_r - \overline{u})^2\right] =$$

$$= \tfrac{b^2}{r}\left[(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \cdots + (z_r - \overline{z})^2\right] = b^2\sigma_Z^2 \ .$$

- *da quanto visto in precedenza, segue allora che la varianza dell'insieme di dati*

$$bz_1 + a \ , \ \ bz_2 + a \ , \ \ \ldots \ , \ \ bz_r + a$$

è uguale a $b^2\sigma_Z^2$.

- *in particolare, indicando con* $\overline{z}$ *e con* σ_Z *la media aritmetica e la deviazione standard dell'insieme di dati*

$$z_1 \ , \quad z_2 \ , \quad \cdots \ , \quad z_r \ ,$$

la varianza dell'insieme di dati

$$\frac{z_1 - \overline{z}}{\sigma_Z} \ , \quad \frac{z_2 - \overline{z}}{\sigma_Z} \ , \quad \cdots \ , \quad \frac{z_r - \overline{z}}{\sigma_Z}$$

è uguale a 1.

L'operazione di passaggio dai dati z_i *ai dati* $\frac{z_i - \overline{z}}{\sigma_Z}$ *si dice* standardizzazione.

In base a tale operazione la media aritmetica dei dati standardizzati risulta nulla e la varianza unitaria.

Dati bidimensionali: covarianza e coefficiente di correlazione

Un caso importante è quello in cui ad ogni unità della popolazione statistica in esame sono associate due variabili X, Y (ad esempio, peso e statura oppure età e reddito e cosi via). In questo caso l'insieme dei dati sarà costituito da delle coppie numeriche

$$(x_1, y_1)\ ,\ \ (x_2, y_2)\ ,\ \ \ldots,\ (x_n, y_n)\ .$$

In generale, per tale insieme di dati non esisterà una legge funzionale precisa che lega X ed Y, tuttavia potrà darsi che, al variare dell'indice i, quando il valore x_i è minore della media aritmetica $\overline{x}$ anche y_i risulta prevalentemente minore di $\overline{y}$ e, viceversa, quando x_i è maggiore della media aritmetica $\overline{x}$ anche y_i tende ad assumere valori maggiori di $\overline{y}$.

In altri casi potrà presentarsi una tendenza di tipo opposto, nel senso che i valori x_i maggiori di $\overline{x}$ prevalentemente si associano con valori y_i minori di $\overline{y}$

e, all'opposto, i valori x_i minori di $\overline{x}$ prevalentemente si associano con valori y_i maggiori di $\overline{y}$.

Infine, un terzo caso è quello in cui non si manifesta nessuna delle due tendenze suddette.

Una misura numerica del modo in cui i valori x_i tendono ad associarsi ai valori y_i è costituita dalla covarianza *di X, Y definita da*

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \ .$$

La covarianza di X e Y è una misura della tendenza di X e Y ad associarsi prevalentemente secondo valori

X	Y	X	Y	
$1.grande$	$grande,$	$piccolo$	$piccolo$	$(\leftrightarrow \sigma_{XY} > 0)$
$2.grande$	$piccolo,$	$piccolo$	$grande$	$(\leftrightarrow \sigma_{XY} < 0)$

dove con $grande$ indichiamo valori di $X > \overline{x}$ e valori di $Y > \overline{y}$ e analogamente per $piccolo$ indichiamo valori di $X < \overline{x}$ e $Y < \overline{y}$. Tipicamente, nel primo caso

la covarianza sarà positiva *e nel secondo* negativa*. Quando si ha* $Cov(X,Y)=0$*, sono assenti entrambe le tendenze suddette e le variabili* X,Y *si dicono* non correlate*. Indicando con* $Var(X)$ *e con* $Var(Y)$ *le varianze dei due insiemi di dati corrispondenti alle variabili* X,Y *e introdotta la variabile* $Z=X+Y$*, i cui valori costituiscono l'insieme di dati*

$$z_1 = x_1 + y_1 \ , \ z_2 = x_2 + y_2 \ , \ \ldots , \ z_n = x_n + y_n \ ,$$

si può verificare che la varianza di Z è data da

$$Var(Z) = Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y) \ .$$

Analogamente, posto $U=X-Y$*, si può verificare che*

$$Var(U) = Var(X-Y) = Var(X)+Var(Y)-2Cov(X,Y) \ .$$

La covarianza soddisfa le seguenti proprietà :

1. $Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\cdot\overline{y} \ ; = \overline{xy} - \overline{x}\cdot\overline{y}$
2. $Cov(X,X) = Var(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \ ;$
3. $Cov(aX+b, cY+d) = acCov(X,Y)$
   dove a,b,c,d *sono delle costanti reali.*

La proprietà 1 dice che la covarianza è data dalla $media\ del\ prodotto\ meno\ il\ prodotto\ delle\ medie.$

In particolare dalla proprietà 3 si ottiene

$$Cov\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho\ .$$

La quantità adimensionale ρ *è una covarianza normalizzata (cioè non dipende dalle unità di misura utilizzate per calcolare i valori di* X *e* Y*) e si chiama* coefficiente di correlazione *di* X, Y*.*

Posto $X' = aX + b, Y' = cY + d$*, con* $ac > 0$*, si può dimostrare che il coefficiente di correlazione di* X', Y' *coincide con quello di* X, Y*. Inoltre, qualunque sia la coppia* X, Y*, per il coefficiente di correlazione* ρ *vale la seguente proprietà*

$$-1\ \leq\ \rho\ \leq\ 1\ .$$

Infatti, in base alla definizione di ρ si ottiene

$$
\begin{aligned}
&Var\left(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\right)= \\
&Var\left(\frac{X}{\sigma_X}\right)+Var\left(\frac{Y}{\sigma_Y}\right)+2Cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right)= \\
&1+1+2\rho=2(1+\rho)\geq 0\ ,
\end{aligned}
$$

e quindi $\rho \geq -1$. Analogamente

$$
\begin{aligned}
&Var\left(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\right)= \\
&Var\left(\frac{X}{\sigma_X}\right)+Var\left(\frac{Y}{\sigma_Y}\right)-2Cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right)= \\
&1+1-2\rho=2(1-\rho)\geq 0\ ,
\end{aligned}
$$

e quindi $\rho \leq 1$.
Infine, si può dimostrare il seguente risultato

$$|\rho| \ = \ 1 \quad \Longleftrightarrow \quad Y = aX + b\ .$$

Infatti, se $Y = aX + b$ segue $Cov(X,Y) = aCov(X,X) = aVar(X)$. Inoltre, $Var(Y) =$

$a^2 Var(X)$ *e quindi* $\sigma_Y = |a|\sigma_X$. *Allora*

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{a\sigma_X^2}{|a|\sigma_X^2} = \frac{a}{|a|} = \begin{cases} +1\,, & a > 0\,; \\ -1\,, & a < 0\,. \end{cases}$$

Viceversa, se $\rho = 1$ *segue*

$$Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2(1-\rho) = 0$$

e quindi

$$\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = cost\,. \tag{1}$$

Allora le coppie (x_i, y_i) *appartengono tutte alla retta di equazione data dalla (1), perciò tra* X *e* Y *esiste una relazione lineare.*
Se invece $\rho = -1$ *si ha*

$$Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 2(1+\rho) = 0$$

e quindi

$$\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = cost\,. \tag{2}$$

Allora le coppie (x_i, y_i) *appartengono tutte alla retta di equazione data dalla (2) e anche in questo caso tra* X *e* Y *esiste una relazione lineare.*

Come mostrato dal precedente risultato, il coefficiente di correlazione esprime una misura della dipendenza lineare *che sussiste tra* X *e* Y*.*

In questo senso quando la nuvola costituita dai dati (x_i, y_i) *è molto addensata intorno a una retta il coefficiente di correlazione* ρ *avrà un valore vicino a* $+1$ *o* -1 *a seconda che il coefficiente angolare della retta sia positivo o negativo.*

Se invece la nuvola di punti è abbastanza rotonda il valore di ρ *sarà vicino a* 0*.*
Osserviamo che se tra X *ed* Y *c'è un legame* non lineare *può risultare* $\rho = 0$*. Un esempio molto semplice è rappresentato dal seguente insieme di dati bidimensionali*

$$(-2, 4)\,, \quad (-1, 1)\,, \quad (0, 0)\,, \quad (1, 1)\,, \quad (2, 4)\,,$$

che soddisfano la relazione $Y = X^2$ *e per i quali, come si può verificare, risulta*

$$\overline{x} = 0, \overline{y} = 2, \sigma_X = \sqrt{2}, \sigma_Y = \sqrt{\frac{14}{5}}\ , \ \ Cov(X, Y) = \rho = 0\ .$$

Rette di regressione

In molte applicazioni tra le variabili X, Y *può sussistere un legame lineare* $Y = aX + b$*, ma a causa di errori di misura nella rilevazione dei dati non si possono determinare* a *e* b *, oppure la dipendenza non è esattamente lineare ma si ritiene che il legame statistico che intercorre tra* X *e* Y *possa essere approssimato con una opportuna funzione lineare del tipo* $Y = aX + b$*.*

Il metodo che si utilizza per scegliere tra le infinite rette quella che meglio approssima la distribuzione di dati bidimensionali risale a Gauss e Legendre ed è noto come metodo dei minimi quadrati*.*

La logica di tale metodo è la seguente: se i punti (x_i, y_i) *appartenessero tutti ad una retta di equazione* $y = ax + b$ *risulterebbe* $(y_i - ax_i - b)^2 = 0$ *per ogni indice* i*. Se una tale retta non esiste, si determina la retta che rende minima la somma dei quadrati, ovvero si determina la coppia* (a, b) *(a cui corrisponde*

la cosiddetta retta di regressione*) per la quale risulta* minima *la quantità*

$$\begin{aligned} f(a,b) &= \sum_{i=1}^{n}(y_i - ax_i - b)^2 \\ &= (y_1 - ax_1 - b)^2 + \cdots + (y_n - ax_n - b)^2. \end{aligned}$$

Calcolando le derivate parziali

$$\begin{cases} \frac{\partial f(a,b)}{\partial a} = -2\sum_{i=1}^{n} x_i(y_i - ax_i - b) \\ \frac{\partial f(a,b)}{\partial b} = -2\sum_{i=1}^{n}(y_i - ax_i - b) \end{cases}$$

e ponendole uguali a zero si ha un sistema che ammette una unica soluzione. Infatti da

$$\begin{cases} \sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = 0 \end{cases}$$

segue

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ n\overline{y} - an\overline{x} - nb = 0 \end{cases}$$

$$\begin{cases} n\overline{xy} - an\overline{x^2} - nb\overline{x} = 0 \\ b = \overline{y} - a\overline{x} \end{cases}$$

$$\begin{cases} n\overline{xy} - an\overline{x^2} - n(\overline{x} \cdot \overline{y} - a\overline{x}^2) = 0 \\ b = \overline{y} - a\overline{x} \end{cases}$$

$$\begin{cases} n\overline{xy} - an\overline{x^2} - n\overline{x} \cdot \overline{y} + an\overline{x}^2 = 0 \\ b = \overline{y} - a\overline{x} \end{cases}$$

$$\begin{cases} n(\overline{xy} - \overline{x} \cdot \overline{y}) = an(\overline{x^2} - \overline{x}^2) \\ b = \overline{y} - a\overline{x} \end{cases}$$

$$\begin{cases} a = \frac{(\overline{xy} - \overline{x} \cdot \overline{y})}{(\overline{x^2} - \overline{x}^2)} = \frac{Cov(X,Y)}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X} \\ \\ b = \overline{y} - \rho \frac{\sigma_Y}{\sigma_X} \overline{x} \end{cases}$$

si ricavano i valori richiesti, cioè

$$\begin{cases} a = \rho\frac{\sigma_Y}{\sigma_X} \\ \\ b = \overline{y} - \rho\frac{\sigma_Y}{\sigma_X}\overline{x} \end{cases}$$

ai quali corrisponde la retta di regressione *(di Y su X) di equazione*

$$y = \overline{y} + \rho\,\frac{\sigma_Y}{\sigma_X}(x - \overline{x})$$

che si può anche porre nella forma

$$\frac{y - \overline{y}}{\sigma_Y} = \rho\,\frac{x - \overline{x}}{\sigma_X}\,. \tag{3}$$

Simmetricamente, l'equazione della retta di regressione di X su Y è

$$x = \overline{x} + \rho\,\frac{\sigma_X}{\sigma_Y}(y - \overline{y})\,,$$

che si può anche scrivere

$$\frac{x-\overline{x}}{\sigma_x} = \rho\,\frac{y-\overline{y}}{\sigma_Y}\ .$$

Osserviamo che le rette di regressione contengono il punto $(\overline{x},\overline{y})$.

Rivediamo il significato di ρ, *calcolando la varianza della differenza tra la variabile statistica normalizzata* $\frac{Y-\overline{y}}{\sigma_Y}$ *e la variabile* $\rho\,\frac{X-\overline{x}}{\sigma_X}$ *stimata tramite la regressione lineare.*

$$Var\left(\frac{Y-\overline{y}}{\sigma_Y}-\rho\,\frac{X-\overline{x}}{\sigma_X}\right) = Var\left(\frac{Y}{\sigma_Y}-\rho\,\frac{X}{\sigma_X}\right) =$$

$$Var\left(\frac{Y}{\sigma_Y}\right)+\rho^2 Var\left(\frac{X}{\sigma_x}\right)-2\rho\,Cov\left(\frac{Y}{\sigma_Y},\frac{X}{\sigma_X}\right) =$$

$$= 1+\rho^2-2\rho^2 = 1-\rho^2.$$

Pertanto si ha

$$Var\left(\frac{Y-\overline{y}}{\sigma_Y}-\rho\,\frac{X-\overline{x}}{\sigma_X}\right) = 0 \Longleftrightarrow \rho = \pm 1.$$

Esempio 7 Nella Tabella 6 sono riportati dei dati bidimensionali; le variabili X, Y rappresentano rispettivamente la statura (in cm), con valori elencati in ordine crescente, e il peso (in Kg) dei 28 ragazzi considerati nell'Esempio 2.
Nella tabella sono riportati le medie aritmetiche $\overline{x}, \overline{y}$, le deviazioni standard σ_X, σ_Y, la covarianza di X, Y, il coefficiente di correlazione ρ, l'equazione della retta di regressione di Y su X e l'equazione della retta di regressione di X su Y.
Il valore $\rho = 0.93$ (prossimo a 1) indica una forte correlazione lineare tra X e Y, come mostrato dal grafico di Figura 2.

Tabella 6: Tabella pesi e Altezze

statura	*peso*
x	*y*
158.0	*45.0*
159.0	*50.8*
159.5	*49.0*
160.0	*49.3*
160.7	*50.0*
161.0	*50.0*
161.5	*50.2*
161.8	*49.7*
162.0	*50.5*
163.0	*51.0*
163.4	*51.0*
163.7	*51.5*
164.0	*51.8*
165.0	*51.3*
165.5	*51.5*
165.8	*51.8*
166.0	*52.0*
166.5	*52.3*
166.8	*53.0*
167.0	*53.3*
167.4	*53.5*
167.8	*53.4*
168.0	*54.0*
170.0	*54.8*
171.4	*55.2*
172.6	*55.5*
173.0	*57.2*
177.5	*56.5*

$\overline{x} = 165.3$
$\overline{y} = 51.97$
$\sigma_X = 4.60$
$\sigma_Y = 2.52$
$Cov(X, Y) = 10.81$
$\rho = 0.93$
retta di regressione di Y su X $y = 0.51x - 32.34$
retta di regressione di X su Y $x = 1.71y + 76.5$

Figura 2: Retta di regressione