CAPITOLO 11
LA STATISTICA

11.1 Origini storiche. Statistica descrittiva

La statistica nasce dall'esigenza di possedere informazioni quantitative su fatti o fenomeni collettivi. Ad esempio, il *census*, nell'antica Roma, consisteva in una rilevazione del numero di cittadini e del loro reddito.
Come disciplina, la statistica (moderna) nasce simbolicamente nel 1662, anno in cui John Graunt pubblica le sue *Observations*. Questo rappresentante della *Scuola degli Aritmetici politici* inglesi si distingue per l'introduzione del metodo empirico induttivo nelle cosiddette scienze sociali, in linea con quanto fatto dai grandi scienziati del suo tempo, da Galileo a Leibniz, passando per Newton.
Parallelamente alla statistica, ma indipendentemente da questa, si sviluppa il *calcolo delle probabilità*, sostanzialmente per rispondere alle questioni più spinose poste dal gioco d'azzardo.
Verso la fine del XIX secolo, con l'evolversi della metodologia statistica, anche la matematica ed il calcolo delle probabilità diventano strumenti essenziali per il compimento degli studi statistici.
Nasce così l'**inferenza statistica**, cioè quell'insieme di tecniche che, sulla base dei risultati ottenuti su un gruppo di osservazioni (il *campione*), permette di trarre delle conclusioni su tutto l'insieme oggetto dello studio (la *popolazione*), espresse in termini di probabilità.
Quando si opera sull'intera popolazione si parla di **statistica descrittiva**: lo studio quantitativo di fenomeni collettivi, finalizzato ad una loro descrizione ed all'indagine della loro natura, in modo da ricercarne le cause e fare delle previsioni.

11.2 Il "linguaggio" della statistica

La definizione di statistica descrittiva che abbiamo dato si fonda sull'applicazione del **metodo induttivo**, che consiste nella generalizzazione di fatti empirici osservati. Si parla di **collettivo statistico** o **popolazione** intendendo con ciò un insieme di **unità statistiche** i, di cui si effettua la rilevazione di uno o più **caratteri**. In seguito indicheremo con P la popolazione, con N l'insieme delle n unità statistiche, mentre con X_j il generico elemento dell'insieme dei k caratteri X. In sintesi:

$$P\begin{cases} i \in N\left(i = 1,...,n\right) \\ X_j \in X\left(j = 1,...,k\right) \end{cases}.$$

L'insieme delle unità statistiche può essere finito o infinito, può riferirsi al totale delle unità esistenti (ad esempio gli abitanti di una nazione) oppure ad una parte, n, detta **campione**.
Se il campione viene scelto in modo *casuale*, il suo studio rientra nel campo di applicazione dell'inferenza statistica.
In generale, per un'efficace applicazione del metodo statistico, occorre una definizione precisa dell'insieme delle unità statistiche sulle quali avverrà l'osservazione del carattere (o dei caratteri) dettata dalle esigenze dell'indagine. Scendiamo adesso nel dettaglio per comprendere meglio il significato del concetto di carattere statistico.
Un carattere statistico si articola in **modalità** (ad esempio, il carattere "sesso" ha le modalità maschio e femmina, "l'età" anni, ecc..). Nel caso del carattere sesso la modalità è *qualitativa*, mentre l'età ha una modalità *quantitativa*.

11.3 Distribuzioni statistiche

Tralasciamo per esigenze di sintesi la descrizione delle fasi in cui si articola un'indagine statistica, e concentriamoci sulle operazioni che possono essere compiute sui dati acquisiti. Una **distribuzione statistica** si forma per effetto della classificazione delle N unità statistiche, in base alla modalità del carattere posseduto da ognuna di esse. Le distribuzioni statistiche possono essere classificate in base alle modalità di uno o più caratteri in *semplici*, *doppie*, o *multiple*.

Distribuzioni statistiche semplici

Quando le distribuzioni statistiche si riferiscono ad un carattere qualitativo si dicono **serie**, quando si riferiscono ad un carattere quantitativo si dicono **seriazioni**. Ambedue le distribuzioni possono essere classificate in base alla modalità: si parla di serie (o seriazioni) di **frequenza** quando sulla

modalità si opera mediante *conteggio* (ad esempio il numero di maschi/femmine di una distribuzione basata sul sesso), e di serie/seriazioni di **intensità** quando sulla modalità si opera mediante *misura* (ad esempio, la quantità di petrolio, espressa in una data unità di misura, estratto da un giacimento in un anno).

Si parla di **frequenza assoluta** come del risultato di un conteggio: se k è il numero di modalità del carattere (nel caso del sesso, maschio = 1, femmina = 2, quindi $k = 2$), il numero di unità statistiche che presentano la modalità i-esima si indica con n_i; la somma delle frequenze assolute coincide con il totale delle unità statistiche: $\sum_{i=1}^{k} n_i = N$.

Spesso è utile ragionare sul *rapporto* tra la frequenza assoluta n_i ed il totale delle unità statistiche: si parla in tal caso di **frequenza relativa** $f_i = \frac{n_i}{N}$, e risulta $\sum_{i=1}^{k} f_i = 1$.

Divisione in classi di una variabile continua

Supponiamo di aver condotto un'indagine su una popolazione, al fine di studiare la distribuzione dell'altezza degli individui che la compongono. La numerosità delle unità statistiche suggerisce di raggrupparle in *classi*, cioè in intervalli di valori, ad esempio da 150 a 155 cm, da 155 a 160 cm, e così via. Una suddivisione di questo tipo pone due questioni importanti:

1) L'esatta definizione delle classi in modo che le singole unità statistiche appartengano ad una ed una sola classe;
2) La scelta del numero e dell'ampiezza delle classi in cui avviene la suddivisione della variabile considerata.

Nel primo caso occorre che le classi della suddivisione siano *contigue* (in modo da non presentare sovrapposizioni o discontinuità), e *continue* (in modo da poter assumere tutti i possibili valori del campo di variazione della variabile).

In definitiva, per le variabili continue è possibile definire le classi in due modi: classi chiuse a sinistra e classi chiuse a destra. Non è chiaramente possibile una suddivisione in classi chiuse, dal momento che ciò farebbe mancare la continuità dei valori assumibili, oppure vi sarebbero sovrapposizioni.

Nel caso di una *variabile discreta* (ad esempio, un'indagine sulla popolazione scolastica in cui le classi sono suddivise per numero di alunni) è ancora possibile una suddivisione della variabile in classi che, stavolta, possono essere *chiuse*, proprio perché costituite da valori discreti.

Occupiamoci adesso della questione del numero e dell'ampiezza delle classi. Si tratta di un problema indeterminato, in quanto possono essere presi in considerazione vari criteri per la formazione delle classi. Di solito si fa in modo che le classi abbiano tutte la stessa ampiezza (criterio dell'*ampiezza costante*), e quindi si pone la questione della determinazione del valore dell'ampiezza: se questo valore è grande si formano poche classi, se è piccolo, al contrario, di classi se ne formano anche troppe.

Esempio 1. Distribuzione statistica semplice (seriazione di frequenza) secondo la statura di $N = 50$ unità statistiche

Statura in classi (in *cm*)	Frequenze assolute	Frequenze relative
$140 \le x < 150$	2	0,04
$150 \le x < 160$	9	0,18
$160 \le x < 170$	17	0,34
$170 \le x < 180$	15	0,30
$180 \le x < 190$	5	0,10
$190 \le x < 200$	2	0,04
Totale	**50**	**1,00**

Distribuzioni statistiche doppie. Tabelle a doppia entrata

Consideriamo il caso in cui si ha a che fare con due caratteri statistici quantitativi (variabili), qualitativi (mutabili), oppure misti. Nel primo caso la tabella si dice *di correlazione*, mentre negli altri due si dice *di contingenza*. La tabella a doppia entrata è uno strumento in cui vengono raccolte le unità statistiche che presentano contemporaneamente una determinata modalità del carattere A, ed una del carattere B.

Esempio 2. Distribuzione statistica doppia secondo la statura ed il peso di $N = 272$ unità statistiche

Peso in classi (in *kg*) Statura in classi (in *cm*)	$50 \le m < 60$	$60 \le m < 70$	$70 \le m < 80$	$80 \le m < 90$	$90 \le m < 100$	**Totale**
$140 \le x < 150$	15	12 21,4%	8	2	0	**37**
$150 \le x < 160$	14 22,6%	20 32,3% 35,7%	17 27,4%	8 12,9%	3 4,8%	**62** **100%**
$160 \le x < 170$	7	14 25,0%	19	11	5	**56**
$170 \le x < 180$	5	9 16,1%	15	21	4	**54**
$180 \le x < 190$	0	1 1,8%	5	14	19	**39**
$190 \le x < 200$	0	0	6	8	10	**24**
Totale	**41**	**56** **100%**	**70**	**64**	**41**	***272***

In generale, dati due caratteri statistici $A = \{a_1, \cdots, a_r\}, B = \{b_1, \cdots, b_s\}$, composti rispettivamente da r e da s modalità (nell'esempio 2 A è la statura e B il peso, quindi $r = 6$ e $s = 5$), la tabella a doppia entrata contiene le *frequenze assolute di associazione* n_{ij} *della modalità i* (del carattere A) *con la modalità j* (del carattere B). Una tabella di questo tipo viene denominata **matrice**, ed ogni elemento corrisponde ad una determinata frequenza assoluta di associazione in cui i è l'indice di *riga* e j è l'indice di *colonna*. Di seguito utilizzeremo la seguente notazione:

$N =$ totale delle unità statistiche del collettivo;

$n_{ij} =$ frequenza di associazione della modalità i-esima di A con la modalità j-esima di B;

$n_{is+1} = \sum_{j=1}^{s} n_{ij}$ totale *marginale* della riga i-esima;

$n_{r+1j} = \sum_{i=1}^{r} n_{ij}$ totale *marginale* della colonna j-esima.

Ovviamente $\sum_{i=1}^{r} n_{is+1} = \sum_{j=1}^{s} n_{r+1j} = N$.

Nelle tabelle a doppia entrata possono essere calcolati tre diversi tipi di frequenze relative:

a) *frequenze relative per riga*: $\frac{n_{i1}}{n_{is+1}}, \cdots, \frac{n_{is}}{n_{is+i}}$; permettono di confrontare le distribuzioni parziali del carattere $B \setminus A = a_i$ (in colore rosso nella tabella sopra).

b) *frequenze relative per colonna*: $\frac{n_{1j}}{n_{r+1j}}, \cdots, \frac{n_{rj}}{n_{r+1j}}$; permettono di confrontare le distribuzioni parziali del carattere $A \setminus B = b_j$ (in colore verde nella tabella sopra).

c) *frequenze relative sul totale N*: $f_{ij} = \frac{n_{ij}}{N}$; permettono di confrontare la diversa incidenza delle frequenze di associazione sul totale della popolazione.

11.4 Le medie

La media aritmetica

Sono date *N* unità statistiche sulle quali sono state rilevate le seguenti modalità: $x_1, x_2, \cdots, x_N$. Si definisce **media aritmetica** il numero $M = \frac{x_1 + x_2 + \cdots + x_N}{N} := \frac{\sum_{i=1}^{n} x_i}{N}$.

Nel caso in cui la stessa modalità si ripeta più volte, alle *k* modalità distinte $x_1, \cdots, x_k$ vengono associate le rispettive frequenze assolute $n_1, \cdots, n_k$, con $\sum_{i=1}^{k} n_i = N$, di conseguenza $M = \frac{\sum_{i=1}^{k} n_i x_i}{N}$ o, in termini di frequenze relative $f_i = \frac{n_i}{N}$, $M = \sum_{i=1}^{k} f_i x_i$.

Può accadere che la media *non* coincida con nessuna delle modalità effettivamente rilevate, in tal caso si parla di *media di conto*.

Esempio. Uno studente riporta alla fine del quadrimestre le seguenti valutazioni in matematica: 6, 6+, 5½, 7, 8-, 7+. Si calcoli la media aritmetica delle 6 valutazioni, con la convenzione che i segni + e – rispettivamente aggiungono e tolgono 0,25 all'intero che li precede.

$M = \frac{6 + 6,25 + 5,5 + 7 + 7,75 + 7,25}{6} = \frac{39,75}{6} = 6,625$ è una *media di conto*, in quanto il suo valore non è mai stato effettivamente ottenuto in nessuna valutazione.

Possiamo affermare che la media aritmetica è un indicatore che opera un'*equipartizione* del carattere, nel senso che rappresenta il valore che avrebbero tutte le modalità se il totale di queste fosse equipartito tra le unità statistiche. Nell'esempio sopra, è come se lo studente avesse sempre riportato come valutazione 6,625 in ognuna delle sei verifiche sostenute.

Nel caso in cui la distribuzione è divisa in classi occorre sostituire ad ogni intervallo un valore che, di solito, coincide con il *valore centrale* della classe.

Esempio. Si vuole calcolare la statura media della distribuzione di 272 studenti appartenenti a tutte le terze classi di un determinato Liceo scientifico, per gli intervalli di altezze riportati in tabella.

Statura in classi (in *cm*)	N. studenti
$140 \le x < 150$	37
$150 \le x < 160$	62
$160 \le x < 170$	56
$170 \le x < 180$	54
$180 \le x < 190$	39
$190 \le x < 200$	24
Totale	*272*

$M = \frac{145 \cdot 37 + 155 \cdot 62 + 165 \cdot 56 + 175 \cdot 54 + 185 \cdot 39 + 195 \cdot 24}{272} = 167,5cm$.

Si parla di *media ponderata* quando si attribuisce *un peso* diverso ad ognuna modalità: $\begin{matrix} Modalità \\ Pesi \end{matrix} \begin{pmatrix} x_1, x_2, \cdots, x_k \\ p_1, p_2, \cdots, p_k \end{pmatrix}$. Si ha $M = \frac{\sum_{i=1}^{k} p_i x_i}{\sum_{i=1}^{k} p_i}$. E' da notare che i pesi svolgono, formalmente, un ruolo analogo a quello delle frequenze assolute, pur avendo un diverso significato.

Esempio. In un mercato, in un determinato giorno, vengono rilevati i seguenti prezzi di agrumi:

Agrumi	Prezzo €/kg	Quantità vendute (kg)
Arance	1,50	100
Limoni	1,60	200
Mandarini	1,00	300

Per calcolare il *prezzo medio* degli agrumi occorre fare una media ponderata:

$$P = \frac{1,50 \cdot 100 + 1,60 \cdot 200 + 1,00 \cdot 300}{100 + 200 + 300} = \frac{770}{600} = 1,28€ \ .$$

In un certo senso, il *centro di massa* di un sistema di N masse poste a varie distanze da un'origine fissata è un esempio di media ponderata: $\vec{r}_{CM} = \sum_{k=0}^{N} \frac{m_k \vec{r}_k}{M}$, dove $M = m_1 + m_2 + m_3 + \ldots + m_N := \sum_{k=0}^{N} m_k$.

Proprietà della media aritmetica

1. Si ha $x_{\min} \le M \le x_{\max}$. Inoltre, se $x_{\min} = M$ allora $x_{\max} = M$ e $x_1 = x_2 = \cdots = x_N = M$.
   - *Dimostrazione*. Si ha $M = \frac{1}{N}\sum_{k=1}^{N} x_k \Rightarrow NM = \sum_{k=1}^{N} x_k \Rightarrow Nx_{\min} \le \sum_{k=1}^{N} x_k = NM \le Nx_{\max}$. Nel caso in cui $x_{\min} = M$, se, per assurdo, $\exists x_j > x_{\min} \Rightarrow M > x_{\min}$; di conseguenza $x_{\max} = M$ e $x_1 = x_2 = \cdots = x_N = M$.
2. La media è il punto di minimo della funzione $f(t) := \frac{1}{N}\sum_{k=1}^{N}(x_k - t)^2$.
   - *Dimostrazione*. Si osserva che la funzione è quadratica nell'incognita *t*, e quindi la sua rappresentazione grafica sul piano (t, y) è una parabola con la concavità rivolta verso l'alto; di conseguenza il minimo viene assunto in corrispondenza del vertice:
     $$f(t) := \frac{1}{N}\sum_{k=1}^{N}(x_k - t)^2 = t^2 - \frac{2t}{N}\sum_{k=1}^{N} x_k + \frac{1}{N}\sum_{k=1}^{N} x_k^2 \Rightarrow t_{\min} = -\frac{b}{2a} = \frac{1}{N}\sum_{k=1}^{N} x_k = M \ .$$
3. La somma degli scarti dalla media è nulla. Di conseguenza, la media può essere definita come quel numero per cui è nulla la somma degli scarti da esso.
   - *Dimostrazione*. $\sum_{k=1}^{N}(x_k - M) = \sum_{k=1}^{N} x_k - NM = NM - NM = 0$.
4. A differenza della somma degli scarti, la *somma dei quadrati degli scarti* è, in generale, diversa da zero, e costituisce un indice della **dispersione** delle modalità intorno alla media. Tale somma si dice **varianza** $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - M)^2$, e gode delle seguenti proprietà:

a) $\sigma^2 \ge 0; \quad \sigma^2 = 0 \Leftrightarrow x_k = M \forall k = 1, \cdots, N$.

b) $\sigma^2 = \left(\frac{1}{N}\sum_{k=1}^{N} x_k^2\right) - M^2$. Infatti,

$$\sigma^2 = \frac{1}{N}\sum_{k=0}^{N} x_k^2 - \frac{2M}{N}\sum_{k=0}^{N} x_k + \frac{1}{N}NM^2 = \frac{1}{N}\sum_{k=0}^{N} x_k^2 - 2M^2 + M^2 = \left(\frac{1}{N}\sum_{k=0}^{N} x_k^2\right) - M^2 \ .$$

c) In caso di modalità ripetute (quindi K modalità distinte), la varianza può essere espressa in termini di frequenze relative $f_i = \frac{n_i}{N}$: $\sigma^2 = \sum_{i=1}^{K} f_i (x_i - M)^2$. Infatti,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - M)^2 = \frac{1}{N}\sum_{i=1}^{K} n_i (x_i - M)^2 = \frac{1}{N}\sum_{i=1}^{K} N f_i (x_i - M)^2 = \sum_{i=1}^{K} f_i (x_i - M)^2 \ .$$

La media geometrica

Consideriamo lo sviluppo $0 \le (x-y)^2 = x^2 + y^2 - 2xy$, dove x, y sono numeri reali positivi. Questo può essere scritto nella forma $xy \le \frac{x^2+y^2}{2}$; si osserva che l'uguaglianza vale solo nel caso in cui $x = y$. Ora, si ha che $xy \le \frac{x^2+y^2+2xy-2xy}{2} = \frac{(x+y)^2}{2} - xy \Rightarrow 2xy \le \frac{(x+y)^2}{2}$, da cui segue $xy \le \left(\frac{x+y}{2}\right)^2 \Rightarrow \sqrt{xy} \le \frac{x+y}{2}$. L'interpretazione geometrica del risultato ottenuto può essere la seguente: *tra tutti i rettangoli di perimetro fissato, quello di area massima è il quadrato* (cioè nel caso in cui $x = y$). Poiché l'espressione $\frac{x+y}{2}$ è la *media aritmetica* delle misure dei lati del rettangolo, si definisce la loro **media geometrica** il numero $M_g = \sqrt{xy}$. In generale, considerata la distribuzione

$$\begin{matrix} \textit{Modalità} \\ \textit{frequenze} \end{matrix} \begin{pmatrix} x_1, x_2, \cdots, x_k \\ n_1, n_2, \cdots, n_k \end{pmatrix},$$ la media geometrica è il numero:

$$M_g = \sqrt[N]{x_1^{n_1} \bullet \cdots \bullet x_k^{n_k}} \quad n_1 + \cdots + n_k = N \quad .$$

Esempio. La somma di 10.000€ viene depositata in banca con un tasso di interesse (*composto*) variabile: 4% il primo anno, 5% il secondo anno, 6% il terzo anno. Qual è il tasso di interesse medio praticato dalla banca?

- Indichiamo con *i* il tasso medio. Per la regola di calcolo dell'interesse composto si ha, sui tre anni, $10.000(1+i)^3 = \{[(10.000(1,04))(1,05)](1,06)\}$. Di conseguenza, $1 + i = \sqrt[3]{(1,04)(1,05)(1,06)} \Rightarrow i = 4,997\%$.

11.5 Moda, mediana, percentili

Spesso è utile disporre di parametri (*descrittori*) al fine di evidenziare alcune caratteristiche dei dati. In particolare, si definiscono *valori modali* quelle modalità la cui frequenza è massima. Se vi è una sola modalità con questa caratteristica, questa si dice **moda** del carattere.

Sono date *N* unità statistiche sulle quali sono state rilevate le seguenti modalità: $x_1, x_2, \cdots, x_N$ riferite ad un carattere quantitativo. Ordiniamo i dati in ordine crescente: $x_{(1)} \le x_{(2)} \le \cdots \le x_{(N)}$.

- Se $N = 2k+1$ si dice **mediana** il valore $x_{med} = x_{\left(\frac{n+1}{2}\right)}$.
- Se $N = 2k$ si dice **mediana** la media aritmetica $x_{med} := \frac{x_{(k)} + x_{(k+1)}}{2}$.

Il significato di mediana è quello di elemento che divide in due parti uguali l'insieme delle modalità rilevate.

Si definisce *k-percentile* quel valore a sinistra del quale si trova il *k*% dei valori rilevati, e a destra l'(1-*k*)%. In base a questa definizione, la mediana è il 50-percentile. Ordinati i valori in senso crescente, il primo quartile ha a sinistra il 25% dei valori, il secondo quartile il 50%, il terzo quartile il 75%.

Interessante è la ripartizione della popolazione in decili. Questa modalità di rappresentazione della distribuzione è estremamente efficace quando si tratta di rappresentare il reddito medio annuo di una famiglia. La tabella che segue ha per fonte la *Banca d'Italia,* supplemento al *"Bollettino statistico"* 2002 (i dati sono riferiti al 2000)

Decimi di reddito	Quote di reddito	Valore di ripartizione (euro)	Quota di famiglie (valori percentuali)	Reddito medio (euro)
Fino al 1° decile	**2,1**	**14.270**	**27,5**	**9.478**
Dal 1° al 2° decile	**4,0**	**19.222**	**15,6**	**16.735**
Dal 2° al 3° decile	**5,2**	**23.323**	**12,3**	**21.224**
Dal 3° al 4° decile	**6,4**	**28.170**	**10,1**	**25.688**
Dal 4° al 5° decile	**7,6**	**32.702**	**8,7**	**30.278**
Dal 5° al 6° decile	**8,9**	**37.908**	**7,4**	**35.222**
Dal 6° al 7° decile	**10,7**	**44.106**	**6,4**	**40.562**
Dal 7° all'8° decile	**12,8**	**53.681**	**5,4**	**48.100**
Dall'8° al 9° decile	**15,7**	**74.746**	**4,2**	**62.202**
Oltre il 9° decile	**26,6**	**-**	**2,4**	**111.072**

Dalla lettura della tabella si osserva che il 10% delle famiglie italiane ha un reddito, nel 2000, inferiore a 14.270€, mentre il 10% ha un reddito superiore a 74.746€.

11.6 La variabilità: lo scarto quadratico medio

Nei fenomeni quantitativi si possono avere diverse distribuzioni che danno la stessa media aritmetica. Ad esempio, lo studente A riporta le seguenti valutazioni in matematica: 6, 7, 7, 8, 7, mentre lo studente B: 5, 5, 8, 8, 9. Ambedue riportano la stessa media: $M_A = \frac{6+7+7+8+7}{5} = 7$, $M_B = \frac{5+5+8+8+9}{5} = 7$. L'attitudine di un fenomeno ad assumere diverse modalità rappresenta la cosiddetta **variabilità**, che viene definita attraverso opportuni *indici* che la misurano. Tra questi il più diffuso è lo **scarto quadratico medio** (*standard deviation*), definito come la radice quadrata della media aritmetica dei quadrati degli scarti:

$$\sigma = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(x_k - M\right)^2}\,.$$

Calcoliamo lo scarto quadratico medio delle distribuzioni dei voti in matematica dei due studenti A e B, al fine di studiarne la dispersione:

$$\sigma_A = \sqrt{\frac{1\cdot(6-7)^2 + 3(7-7)^2 + 1\cdot(8-7)^2}{5}} = 0,632$$
$$\sigma_B = \sqrt{\frac{2\cdot(5-7)^2 + 2(8-7)^2 + 1\cdot(9-7)^2}{5}} = 1,673\,.$$

Possiamo concludere che le modalità dello studente B sono più *disperse* intorno alla media. L'importanza dello scarto quadratico medio è legata anche al fatto che, insieme alla media aritmetica, caratterizza la cosiddetta *distribuzione normale* o *di Gauss*, utilizzata in fisica nella trattazione degli errori casuali.

Requisiti formali di un indice di variabilità

Un indice di variabilità (d'ora in poi, per noi, lo scarto quadratico σ) deve possedere i seguenti requisiti:

- Deve essere nullo se e solo se le modalità del carattere sono tutte uguali, positivo altrimenti;
- Deve essere invariante per moltiplicazione di tutte le frequenze per una costante positiva:

$$\begin{pmatrix} x_1, x_2, \cdots, x_k \\ An_1, An_2, \cdots, An_k \end{pmatrix} \Rightarrow \sigma' = \sqrt{\frac{An_1\left(x_1 - M\right)^2 + \cdots + An_k\left(x_k - M\right)^2}{\left(An_1 + \cdots + An_k\right)}} = \sigma\,;$$

- Deve coincidere per due distribuzioni in cui le modalità differiscono per una costante. Ciò significa, ad esempio, che la variabilità della temperatura non deve dipendere dalla particolare scala utilizzata per la sua misura:
$$\begin{pmatrix} x_1+B, x_2+B, \cdots, x_k+B \\ n_1, n_2, \cdots, n_k \end{pmatrix} \Rightarrow \sigma' = \sqrt{\frac{n_1\left(x_1+B-M-B\right)^2+\cdots+n_k\left(x_k+B-M-B\right)^2}{\left(n_1+\cdots+n_k\right)}} = \sigma .$$
- Deve essere espresso nella stessa unità di misura delle modalità.

Proprietà dello scarto quadratico medio

Abbiamo già avuto modo di vedere che $\sigma^2 = \left(\frac{1}{N}\sum_{k=1}^{N} x_k^2\right) - M^2$, da cui segue

$\sigma = \sqrt{\left(\frac{1}{N}\sum_{k=1}^{N} x_k^2\right) - M^2}$.

Esercizio. In relazione alla seconda prova scritta dell'Esame di Stato 2103 (Liceo scientifico di ordinamento) sono state rilevate, in una commissione composta da due classi, le seguenti valutazioni. Si calcoli la media dei voti e lo scarto quadratico medio.

Modalità (voto in 15-esimi)	frequenze
5	2
6	1
7	4
8	4
9	7
10	2
11	2
12	3
13	6
14	5
15	6

11.7 Disuguaglianza di Bienaymé-Cebicev

Si vuol disporre di una stima della percentuale di unità statistiche che presentano modalità che differiscono dalla media *più* di un valore fissato, *t*.

Teorema. Sia $\left(x_1, \cdots, x_N\right)$ il risultato della rilevazione di un carattere *x*, con media M e varianza σ^2, su una popolazione di N unità statistiche. Per $t \geq 1$, sia $S_t := \left\{ x_i \,\middle|\, i = 1, \cdots, N; \quad \left|x_i - M\right| > t \right\}$[1]. Allora, la percentuale delle modalità che differiscono dalla media più di *t*, $\frac{\#S_t}{N}$, è tale che $\frac{\#S_t}{N} \leq \frac{\sigma^2}{t^2}$.

Dimostrazione. Dalla definizione di varianza, segue $N\sigma^2 = \sum_{i=1}^{N}\left(x_i - M\right)^2$. Poiché $\bigcup_{t\geq 1} S_t = \left(x_1, \cdots, x_N\right)$, per *t* fissato l'insieme delle modalità si può dividere in due sottoinsiemi: S_t ed il suo complementare in $\left(x_1, \cdots, x_N\right)$. Così facendo, risulta

$N\sigma^2 = \sum_{i=1}^{N}\left(x_i - M\right)^2 = \sum_{x_i \in S_t}\left(x_i - M\right)^2 + \sum_{x_i \notin S_t}\left(x_i - M\right)^2 \geq \sum_{x_i \in S_t}\left(x_i - M\right)^2 > \sum_{x_i \in S_t} t^2 = \#S_t \cdot t^2$. In definitiva, quindi, $\frac{\#S_t}{N} \leq \frac{\sigma^2}{t^2}$.

---

[1] L'insieme delle modalità che differiscono dalla media più di un valore fissato *t*.

Se volessimo valutare la percentuale di modalità che differiscono dalla media *meno* di un valore fissato *t*, dovremmo ragionare sul complementare dell'insieme S_t. Posto

$\overline{S}_t := \left\{ x_i \;\middle|\; i = 1, \cdots, N; \quad |x_i - M| < t \right\}$, allora $\frac{\#\overline{S}_t}{N} \geq 1 - \frac{\sigma^2}{t^2}$. In particolare, se $t = k\sigma$, l'insieme

$\overline{S}_t := \left\{ x_i \;\middle|\; i = 1, \cdots, N; \quad |x_i - M| < t \right\}$ rappresenta l'insieme delle modalità comprese nell'intervallo delimitato dalla media più o meno *k* volte lo scarto quadratico; la frazione di modalità comprese in questo insieme è $\frac{\#\overline{S}_{k\sigma}}{N} \geq 1 - \frac{1}{k^2}$. Il teorema di cui sopra, espresso in questi termini, è noto come *Disuguaglianza di Bienaymé-Cebicev*.

Esempio. Supponiamo che la media dei redditi dichiarati da un gruppo di famiglie sia di 25.000 euro in un determinato anno, con uno scarto quadratico medio pari a 5.000 euro. Una famiglia si può considerare *benestante* se dichiara un reddito non inferiore a 40.000 euro. Qual è la percentuale di famiglie benestanti?

Utilizzando la disuguaglianza del teorema (di Cebicev) con $t = 40.000 - 25.000 = 15.000$ si ha

$\frac{\#S_t}{N} \leq \frac{\sigma^2}{t^2} \Rightarrow \frac{\#S_t}{N} \leq \left(\frac{5.000}{15.000} \right)^2 = 11{,}11\%$.

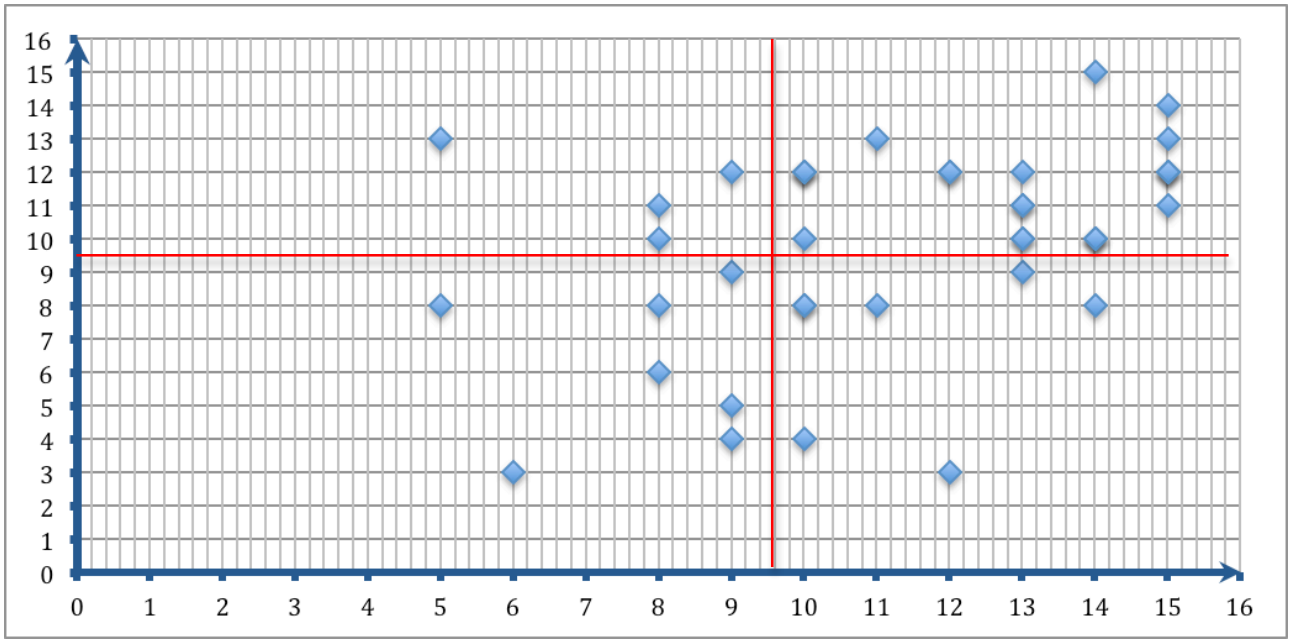
E' rilevante poter disporre di una stima su una popolazione senza conoscere, in linea di principio, di quante unità statistiche si compone.

11.8 Relazioni statistiche

Supponiamo di voler rilevare *due* caratteri sulla stessa unità statistica, e di voler indagare alla ricerca di una relazione tra questi; in altre parole, si tratta di cercare una **connessione** tra la variazione del carattere X, e quella del carattere Y al variare di uno dei due.
Supponiamo di aver rivelato, su una popolazione di 40 studenti che hanno affrontato l'esame di Stato, il voto riportato nella seconda prova scritta (Matematica) e nella terza prova (Scienze).

Studente	Voto Matematica (in 15-esimi)	Voto Scienze (in 15-esimi)
1	6	3
2	10	12
3	14	10
4	15	13
5	14	10
6	10	12
7	8	10
8	10	10
9	10	8
10	9	4
11	12	12
12	15	12
13	11	13
14	10	8
15	13	9
16	9	5
17	14	8
18	9	9
19	13	11
20	13	11
21	15	14
22	13	10
23	14	10
24	15	12
25	10	12
26	5	8
27	5	13
28	13	12
29	9	12
30	10	4
31	11	8
32	12	12
33	15	12
34	14	15
35	8	6
36	15	11
37	8	11
38	8	8
39	12	3
40	13	10

Riportiamo su un grafico la distribuzione delle valutazioni, matematica sull'asse delle ascisse, scienze sull'asse delle ordinate.



Dal grafico non si può evidenziare una chiara connessione tra l'esito della prova di matematica e quella di scienze.
Poiché non è molto realistico supporre che uno studente riporti la stessa valutazione nelle due prove, è opportuno organizzare in *classi* la distribuzione. Ad esempio, riportiamo nella prima colonna le modalità riferite alle valutazioni in matematica, e nella prima riga quelle riferite a scienze. Gli elementi della tabella rappresentano le frequenze associate alla modalità *ij*, dove *i* è il voto in matematica e *j* il voto in scienze.

	<8	**8-9**	**10-11**	**12-13**	**14-15**
<8	1	1	0	1	0
8-9	3	2	2	1	0
10-11	1	3	1	4	0
12-13	1	1	4	3	0
14-15	0	1	4	4	2

Non ci soffermeremo sulle tecniche d'indagine utilizzate nelle tabelle a doppia entrata, bensì concentreremo i nostri sforzi sulla ricerca di una relazione funzionale tra le due variabili statistiche, basandoci sull'esame della distribuzione delle modalità di cui disponiamo.

11.9 Il metodo dei minimi quadrati

In fisica si ha spesso a che fare con esperimenti che richiedono la misura di molti valori di *due* grandezze fisiche, con lo scopo di individuare la relazione matematica che le lega. Cominciamo il nostro ragionamento con la **relazione lineare** tra due grandezze fisiche. Possiamo individuare

due questioni rilevanti: la ricerca della retta che si adatta meglio alla distribuzione di misure (la cosiddetta *regressione lineare*[2]), e la determinazione di *quanto è buono* l'adattamento delle misure rilevate alla retta trovata. Per la seconda questione occorre distinguere il caso in cui conosciamo le incertezze nelle misure da quello in cui non le conosciamo. Nel primo caso, infatti, la rappresentazione delle misure con le *barre d'errore* è strategica nella ricerca *grafica* della retta, cosiddetta di *best-fit*.
Occorre precisare che, in generale, i punti del piano corrispondenti alle coppie di misure non appartengono alla retta di regressione individuata analiticamente, o alla retta di best-fit.
Il metodo utilizzato nella ricerca dei coefficienti della retta di regressione è il cosiddetto **metodo dei minimi quadrati**. Vediamo in cosa consiste.
Indicate con (x_i, y_i) le coordinate di N punti del piano, vogliamo determinare i coefficienti della retta $y = mx + q$ che meglio approssima la distribuzione di punti, ovvero la retta per la quale è minima la quantità

$$\varepsilon = \sum_{i=1}^{N} (mx_i + q - y_i)^2$$ (per questo motivo si parla di minimi quadrati).

La funzione introdotta, che dipende dalle variabili m, q, fornisce una misura dell'errore totale che si ottiene sostituendo i punti *reali* della distribuzione, con quelli *teorici* corrispondenti sulla retta.
Per determinare la coppia m, q che minimizza la funzione $\varepsilon(m,q)$ conviene ragionare *separatamente*, ovvero fissando uno dei due parametri e ragionando in funzione dell'altro, notando che, in ambedue i casi, la funzione è *quadratica* nel parametro considerato:

$$\varepsilon(m) = m^2 \sum_{i=1}^{N} x_i^2 + 2m \sum_{i=1}^{N} x_i (q - y_i) + \sum_{i=1}^{N} (q - y_i)^2 = am^2 + bm + c \Rightarrow m_{\min} = -\frac{b}{2a}$$, oppure

$$\varepsilon(q) = Nq^2 + 2q \sum_{i=1}^{N} (mx_i - y_i) + \sum_{i=1}^{N} (mx_i - y_i)^2 = aq^2 + bq + c \Rightarrow q_{\min} = -\frac{b}{2a}$$.

Ricordando che le funzioni quadratiche (corrispondenti a parabole con la concavità rivolta verso l'alto) assumono il minimo in corrispondenza del vertice della parabola, i valori m, q che minimizzano la funzione $\varepsilon(m,q)$ sono le soluzioni del sistema formato dalle condizioni $m_{\min} = -\frac{b}{2a}$ e $q_{\min} = -\frac{b}{2a}$:

$$\begin{cases} m = -\dfrac{b}{2a} = -\dfrac{2\sum_{i=1}^{N} x_i (q - y_i)}{2\sum_{i=1}^{N} x_i^2} \\ q = -\dfrac{b}{2a} = -\dfrac{2\sum_{i=1}^{N} (mx_i - y_i)}{2N} \end{cases} \Rightarrow \begin{cases} m\sum_{i=1}^{N} x_i^2 + q\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i y_i \\ m\sum_{i=1}^{N} x_i + Nq = \sum_{i=1}^{N} y_i \end{cases}$$. Le soluzioni del sistema sono quindi:

[2] Il termine "regressione" fu introdotto dallo statistico inglese Francis Galton (1886), il quale osservò che le stature dei figli di padri alti (cioè con statura superiore alla media) tendevano ad avvicinarsi alla statura media, in altre parole a regredire verso la media.

$$m = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}}{N\left(\frac{\sum_{i=1}^{N} x_i^2}{N} - \overline{x}^2\right)} .$$

$$q = \frac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

Il metodo dei minimi quadrati vale per *tutte* le relazioni analitiche tra due variabili, ma per dimostrare ciò occorre impiegare strumenti più avanzati di quelli utilizzati nel caso della relazione lineare. Tuttavia, esistono delle relazioni che possono essere *linearizzate* attraverso semplici trasformazioni.

La relazione esponenziale $y = ab^x$

In questo caso, una trasformazione logaritmica in base *e* permette di scrivere la relazione esponenziale nella forma:

$$\log y = \log a + x \log b .$$

Posto $z := \log y$, $q := \log a$, e $m := \log b$, la relazione trasformata può essere scritta nella forma

$$z = mx + q ,$$

e a questa possiamo applicare il metodo dei minimi quadrati per calcolare i valori m, q. I valori cercati a, b si determinano dalle relazioni:

$$\begin{aligned} a &= e^q \\ b &= e^m \end{aligned} .$$

11.10 Covarianza

Si definisce **covarianza** la quantità $\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{N}$.

Proposizione. $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}$.

Dimostrazione. $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{N} x_i y_i - \overline{y}\sum_{i=1}^{N} x_i - \overline{x}\sum_{i=1}^{N} y_i - \sum_{i=1}^{N}\overline{x}\cdot\overline{y} =$
$\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y} - N\overline{x}\cdot\overline{y} + N\overline{x}\cdot\overline{y} = \sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}$.

In particolare, ricordando la definizione di varianza, il coefficiente angolare della retta approssimante può quindi essere scritto nella forma:

$$m = \frac{\sigma_{xy}}{\sigma_x^2} .$$

Esercizio. Dimostrare la seguente proprietà della covarianza: $|\sigma_{xy}| \le \sigma_x^2 \cdot \sigma_y^2$.

Suggerimento: si interpreti la quantità $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$ come il prodotto scalare di due vettori opportuni…

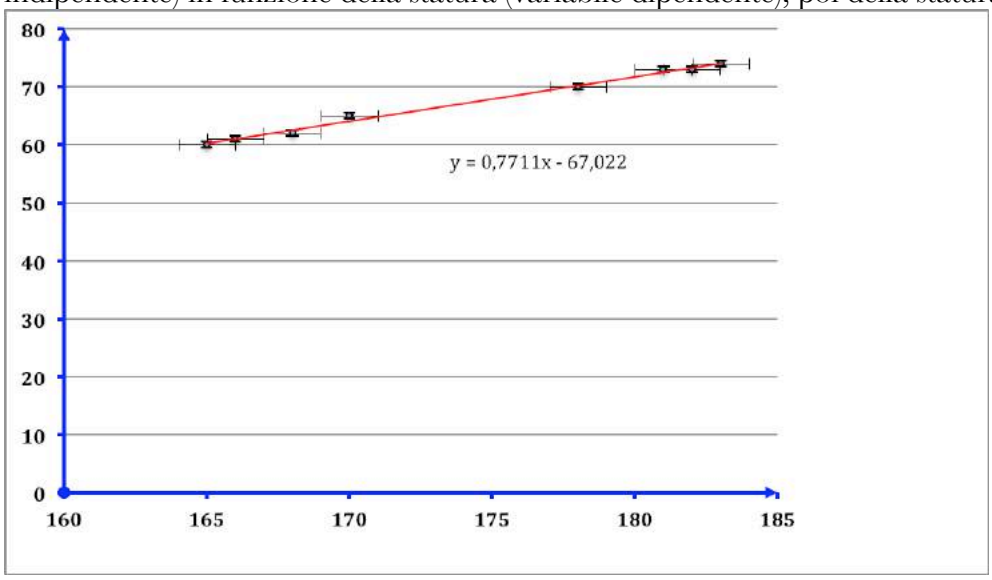
11.11 Il coefficiente di correlazione lineare

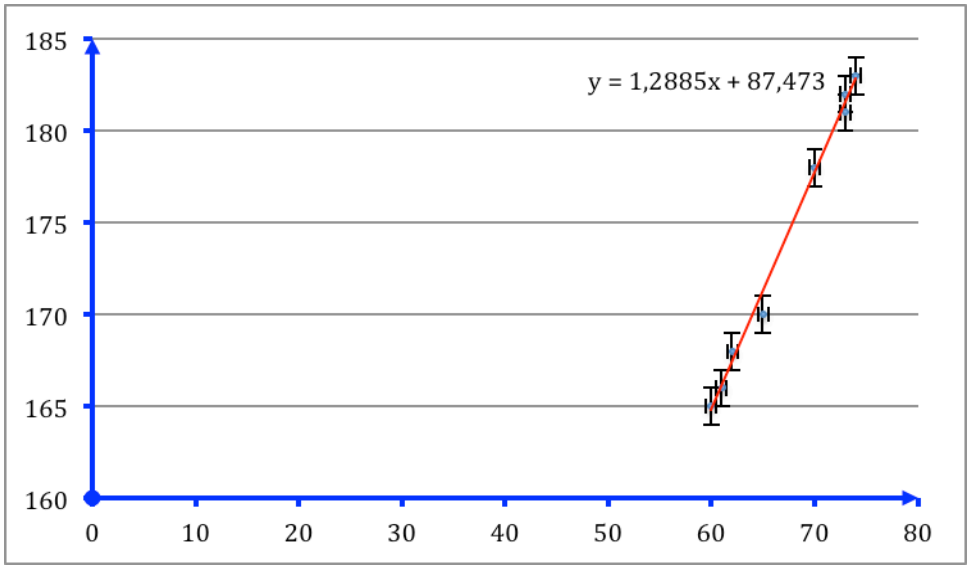
In fisica si dispone quasi sempre di una stima affidabile delle incertezze, quindi l'operazione di best-fit da sola può essere sufficiente per stabilire il tipo di relazione tra le variabili con una ragionevole sicurezza. Tuttavia, specialmente nelle scienze sociali, di solito non si dispone di alcuna incertezza

sui dati rilevati, e la *correlazione* (lineare) deve essere stabilita in un altro modo; occorre quindi stabilire un criterio di valutazione opportuno. Esaminiamo il seguente esempio.
Esempio. Supponiamo di aver misurato il peso e la statura di 8 allievi di una società di atletica leggera, intenti a partecipare ad una gara di velocità (tratto dal testo *Statistica* Fraire – Rizzi).

Allievo	Peso in kg	Statura in cm
1	70	178
2	60	165
3	73	181
4	61	166
5	65	170
6	62	168
7	73	182
8	74	183

Riportiamo i dati su un grafico e cerchiamo una relazione di dipendenza, prima del peso (variabile indipendente) in funzione della statura (variabile dipendente), poi della statura in funzione del peso.





Dai due grafici appare piuttosto evidente la relazione di linearità tra le due variabili. I coefficienti lineari delle rette di regressione (calcolati con il metodo dei minimi quadrati) sono rispettivamente $m = 0,771$ e $m' = 1,289$. Ora, se scambiamo tra loro le variabili al fine di verificarne l'interdipendenza, ci aspettiamo che i coefficienti angolari siano l'uno il *reciproco* dell'altro: $m' = \frac{\sigma_{xy}}{\sigma_y^2} \approx \frac{1}{m} = \frac{\sigma_x^2}{\sigma_{xy}}$. Da questo fatto segue in modo alquanto naturale la definizione di

coefficiente di correlazione lineare di Bravais-Pearson:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y},$$

in base al quale due variabili sono tanto più correlate quanto il suo valore si avvicina a 1 (o a -1, se al crescere dell'una diminuisce l'altra). Valori di $r \approx 0$ indicano una non correlazione (lineare!) tra le due variabili.

Di seguito viene riportato lo studio condotto mediante utilizzo del foglio elettronico excel.

Allievo	Statura in cm	Peso in kg	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$	Allievo	Peso in kg	Statura in cm	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
1	178	70	12460	31684	4900	1	70	178	12460	4900	31684
2	165	60	9900	27225	3600	2	60	165	9900	3600	27225
3	181	73	13213	32761	5329	3	73	181	13213	5329	32761
4	166	61	10126	27556	3721	4	61	166	10126	3721	27556
5	170	65	11050	28900	4225	5	65	170	11050	4225	28900
6	168	62	10416	28224	3844	6	62	168	10416	3844	28224
7	182	73	13286	33124	5329	7	73	182	13286	5329	33124
8	183	74	13542	33489	5476	8	74	183	13542	5476	33489
	$\bar{x}$	$\bar{y}$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$		$\bar{x}$	$\bar{y}$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
	174,125	67,25	93993	242963	36424		67,25	174,125	93993	36424	242963

σ_x	7,13157591
σ_y	5,51701912
σ_{xy}	39,21875
r=	0,99679018
m=	0,77112135

m*m= 0,99359065

σ_x	5,51701912
σ_y	7,13157591
σ_{xy}	39,21875
m=	1,28850103