I.T.I. "Modesto PANETTI" – B A R I

Via Re David, 186 - 70125 BARI ☎ 080-542.54.12 - Fax 080-542.64.32

Intranet http://10.0.0.222 - Internet http://www.itispanetti.it – email : BATF05000C@istruzione.it

Analisi statistica – Fenomeni aleatori

Dispensa per gli alunni della classe 3ETB

Prof. Ettore Panella

Introduzione

Un sistema si dice **deterministico** se è possibile descriverne il comportamento mediante rigorose formule matematiche. È questo il caso del moto dei pianeti intorno al sole descritto dalle note equazioni della gravitazione di Newton mediante le quali si può determinare l'esatta posizione del pianeta nel tempo.

Un sistema si dice **probabilistico** o stocastico o aleatorio se non è possibile prevedere con certezza i suoi stati futuri ma se ne può solo valutare la probabilità. Ad esempio nel gioco del lotto non è possibile prevedere con certezza i numeri estratti ma si può solo calcolare la probabilità che si verifichi una certa estrazione.

Lo studio dei sistemi stocastici deve, pertanto, utilizzare opportune procedure matematiche atte a fornire i necessari elementi di valutazione del sistema.

La teoria dei giochi, lo studio degli errori nelle misure e l'analisi statistica sulle popolazioni sono esempi di sistemi stocastici.

In questa breve nota si descrivono i metodi fondamentali che sono alla base dello studio dei sistemi stocastici. Per approfondimenti si rimanda a testi specialistici.

Probabilità e Frequenza

Si definisce **probabilità di un evento p** il rapporto tra il numero dei casi favorevoli $\mathbf{N_F}$ e il numero dei casi possibili $\mathbf{N_P}$:

$$p = \frac{N_F}{N_P} \qquad (1)$$

Ad esempio, nel lancio di una moneta con due facce, testa e croce, la probabilità che esca testa si calcola tenendo conto che: $N_F = 1$ e $N_P = 2$ per cui:

$$p = \frac{1}{2} = 0.5$$

In percentuale p = 50 %

La probabilità è sempre un numero minore di 1. Probabilità p = 1 equivale alla certezza.

Si definisce **frequenza di un evento f** il rapporto tra il numero di volte che l'evento si verifica $\mathbf{N_V}$ e il numero totale di prove effettuate $\mathbf{N_T}$.

$$f = \frac{N_V}{N_T}$$

Teorema di Bernulli o ***Legge dei grandi numeri***. Il valore della frequenza tende a quello della probabilità al tendere del numero delle prove all'infinito.

$$p = \lim_{N_P \to \infty} f$$

Ad esempio, se si lancia una moneta 100 volte si misura che si è ottenuta testa 49 volte. Si ha:

$$f = \frac{N_V}{N_T} = \frac{49}{100} = 0.49$$

In percentuale f = 49 % e non 50 %.
Effettuando un numero di prove molto più elevato si può ricavare che $f \to 50\%$.

Distribuzione di probabilità

Si definisce distribuzione di probabilità il grafico della probabilità in funzione del valore assunto dalla variabile aleatoria. Ad esempio in un lancio di un dado l'evento è l'estrazione di un numero compreso tra 1 e 6 e la relativa probabilità è p = 1/6. In fig. 1 si riporta l'istogramma relativo alla distribuzione di probabilità. Tale distribuzione è detta **uniforme** poiché la probabilità è la stessa per tutti gli eventi.

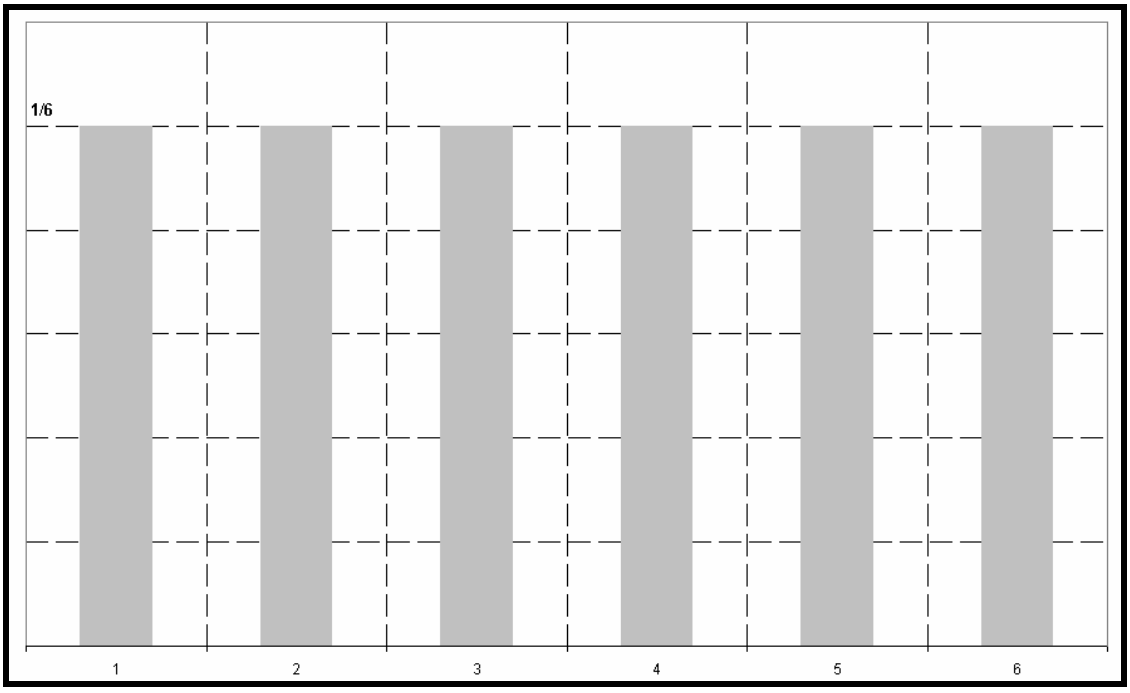


Fig. 1 - Distribuzione di probabilità per il lancio di un dado.

Si consideri il lancio di 2 dadi e l'evento di cui si vuole studiare la probabilità sia quello relativo al calcolo della somma dei numeri usciti su ciascun dado. Tale somma può assumere tutti i valori interi compresi tra 2 e 12. Nella seguente tabella si riportano le possibili estrazioni.

		DADO 1					
		1	**2**	**3**	**4**	**5**	**6**
DADO 2	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Appare evidente che gli eventi possibili sono 36 per cui la probabilità associata a ciascun evento si deve calcolare applicando la definizione (1). Ad esempio l'evento relativo alla somma pari a 12 si verifica solo se ciascun dado fornisce il valore 6. Il numero di casi favorevoli è 1 per cui la relativa probabilità vale: p(12) = 1/36.

La somma uguale a 7 si può ottenere in 6 situazioni diverse per cui la relativa probabilità vale: p(7) = 6/36. In fig. 2 si mostra la distribuzione di probabilità per questo esempio.

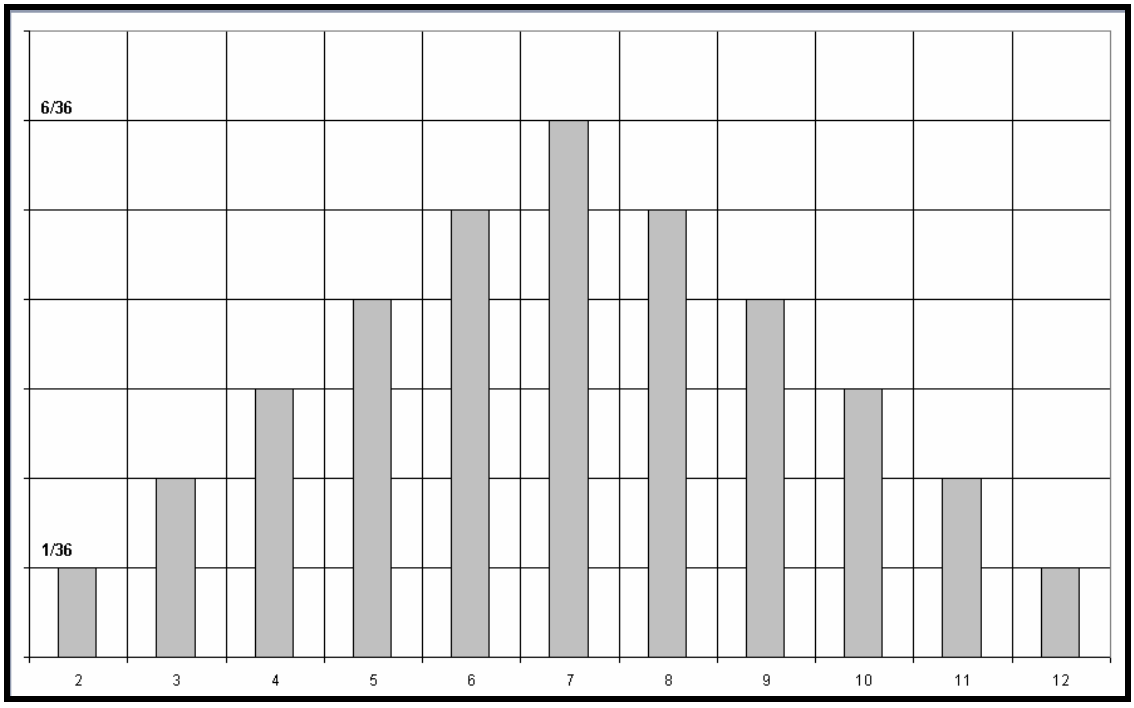


Fig. 2. - Distribuzione di probabilità per il lancio di 2 dadi.

In questo caso la distribuzione di probabilità è **non uniforme** poiché gli eventi non sono tutti equiprobabili.

In tutti i casi è facile verificare che la somma delle probabilità deve valere 1 (evento certo).

$$\sum_{i=1}^{N} p_i = 1$$

Le distribuzioni di probabilità descritte precedentemente si riferiscono a **sistemi discreti** poiché la variabile aleatoria può assumere un numero di valori finito e la relativa rappresentazione grafica è un istogramma.

Esistono sistemi nei quali la variabile aleatoria può assumere infiniti valori all'interno di un determinato intervallo. Si pensi alla misura di una grandezza fisica come la tensione di uscita di un

amplificatore. In questo caso ripetendo la misura più volte si possono trovare innumerevoli valori dovuti agli inevitabili errori sistematici e casuali.

Nei **sistemi continui** la distribuzione di probabilità è una curva continua e viene definita come curva della **densità di probabilità**.

In fig. 3 si mostra il grafico della densità di probabilità per variabile aleatoria continua secondo la **distribuzione di Gauss**. Tale distribuzione è nota anche come **distribuzione normale**.

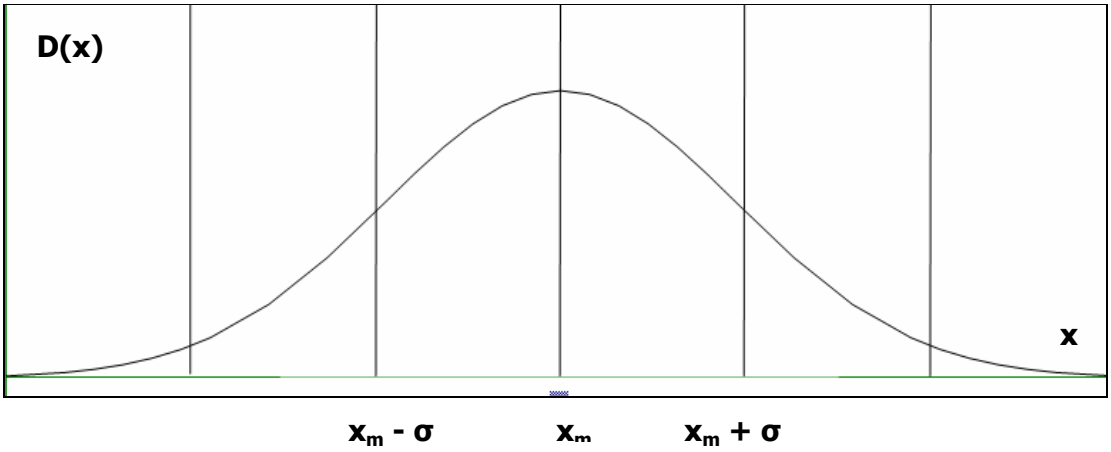


Fig. 3. - Distribuzione di Gauss.

La curva di Gauss ha una caratteristica forma a campana e l'espressione matematica che la descrive vale:

$$D(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-x_m)^2}{2\cdot\sigma^2}}$$

I termini $\mathbf{x_m}$ e $\boldsymbol{\sigma}$ sono denominati rispettivamente **valore medio** e **deviazione standard (standard deviation).** Il quadrato della deviazione standard è noto come **varianza**.

Da quanto detto appare evidente che l'area sottesa dalla curva di Gauss rappresenta la probabilità totale associata al fenomeno aleatorio e pertanto deve valere 1.

La distribuzione di Gauss riveste una notevole importanza pratica poiché numerosi fenomeni fisici di tipo probabilistico seguono tale distribuzione. A tale categoria appartengono gli errori sperimentali nelle misurazioni di grandezze fisiche.

Media e Deviazione Standard

Quando si esegue una misurazione di una qualunque grandezza fisica si commettono inevitabilmente degli errori. Tali errori vengono, normalmente, classificati in :

- **Errori sistematici:** hanno sempre uguale ampiezza e segno e sono dovuti a strumenti di misura alterati, approssimazioni introdotte in conseguenza a relazioni che legano i parametri da misurare, inserzioni di strumenti ecc;
- **Errori accidentali:** hanno ampiezza e segno variabili e sono dovuti a variazioni ambientali ed atmosferiche, vibrazioni meccaniche indesiderate durante la lettura dei dati, imperizia dello sperimentatore (errore di parallasse), attriti meccanici, campi magnetici ed elettrici esterni, ecc. La somma algebrica di questi errori tende a zero se la misura viene ripetuta molte volte;
- **Errori di insensibilità:** sono causati dalla incapacità dello sperimentatore e degli strumenti di misura di apprezzare variazioni inferiori ad un certo limite.

Gli errori sistematici si possono, in parte, compensare se si conosce la classe di precisione degli strumenti e se si pone cura nella esecuzione delle misure.

Gli errori accidentali, invece, non si possono valutare a priori dato il loro carattere del tutto aleatorio. Per compensare in parte tali errori si ricorre alla misura ripetuta della grandezza fisica.

Indicando con:

$$x_1, x_2, \ldots\ldots\ldots\ldots\ldots x_N$$

N misure su una grandezza fisica, con N sufficientemente grande, si definisce **valore medio** e si indica con x_m o con μ o con $\bar{x}$ la quantità:

$$x_m = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots\ldots\ldots\ldots\ldots x_N}{N}$$

Si dimostra che il **valore medio** rappresenta la **migliore stima** della misura nel senso che esso pur non rappresentando il **valore vero** della grandezza (impossibile da trovare a causa degli errori) è senza dubbio quello più probabile.

Si definisce **scarto dalla media s_i** per una generica misura $\mathbf{x}_i$ la quantità:

$$s_i = x_i - x_m$$

Ovviamente gli scarti possono assumere valori sia positivi che negativi. Lo scarto dalla media è noto anche come **errore assoluto** della misura. Si definisce **errore relativo ε_r%** :

$$\varepsilon_r\% = \frac{x_i - x_m}{x_m} \cdot 100$$

La definizione di scarto dalla media consente di introdurre il concetto di **standard deviation σ** che rappresenta l'incertezza media delle misure. Tale parametro si valuta con la seguente formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - x_m)^2}{N-1}}$$

Si considerano gli scarti al quadrato per non tener conto del loro segno, mentre il termine N-1 a denominatore si giustifica tenendo conto che nel caso di una sola misurazione N=1 si ha una forma indeterminata per cui non è possibile, come è giusto che sia, valutare σ**.**

Si vuole sottolineare che la teoria che si sta sviluppando è valida e fornisce buoni risultati solo nel caso di numerose misurazioni.

Si può dimostrare che la probabilità p che una misurazione x sia compresa tra:

- $x_m - \sigma < x < x_m + \sigma$: vale p = 68.27 %
- $x_m - 2\sigma < x < x_m + 2\sigma$: vale p = 95.45 %
- $x_m - 3\sigma < x < x_m + 3\sigma$: vale p = 99.73 %.

In altre parole se le misure sono distribuite secondo la curva di Gauss e si ripetono un numero elevatissimo di volte allora circa il 68 % di tali misure si addensa nell'intervallo $x_m \pm \sigma$.

Errore quadratico medio della media o Deviazione Standard della media

Se si effettuano N misure di una determinata grandezza fisica si è detto che il valore medio rappresenta il valore più probabile rispetto al valore vero e che la standard deviation caratterizza l'incertezza media delle singole misure.

Si può dimostrare che l'incertezza media sul risultato finale x_m risulta essere pari alla deviazione standard divisa per la radice quadrata di N. In formule:

$$\sigma_{x_m} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_m)^2}{N(N-1)}}$$

Tale parametro è noto con molti nomi come **errore standard** o **errore standard della media** o **errore quadratico medio della media** o **deviazione standard della media.**

Ad esempio supponiamo di eseguire un certo numero di misure di una tensione e di ottenere i seguenti valori:

$$x_m = 15.3 \text{ e } \sigma_{x_m} = 0.2$$

Ne segue che il valore della tensione è compreso tra:

$x = 15.3 \pm 0.2$ con una probabilità di circa il 68 %

Oppure che il valore della tensione è compreso tra:

$x = 15.3 \pm 0.4$ con una probabilità di circa il 95 %

Mentre è compreso tra:

$x = 15.3 \pm 0.6$ con una probabilità di circa il 99 %

che rappresenta la quasi certezza.

È chiaro che se l'errore quadratico medio della media è piccolo vuol dire che si sono effettuate delle buone misurazioni con piccoli errori accidentali. In tal caso la curva di Gauss risulta essere molto stretta.

In conclusione si può affermare che in una serie di misure di una variabile aleatoria in cui prevale l'errore accidentale il risultato della misura si può ritenere pari al valore medio con una grado di incertezza desumibile dalla conoscenza dell'errore quadratico medio della media.

Cifre significative

Il numero di cifre significative fornisce una indicazione approssimata dell'errore relativo. Ad esempio, consideriamo i numeri:

120 e 950

Entrambi sono accurati a 2 cifre significative. Per cui si ha:

120 ± 5 all'incirca si può scrivere: 120 ± 5 %

Per il numero 950, sempre con 2 cifre significative, si ha:

950 ± 5 all'incirca si può scrivere: 950 ± 0.5%

Si ricava che l'errore relativo associato a 2 cifre significative varia tra lo 0.5% e il 5% a seconda del valore della cifra più significativa.

Con analogo ragionamento si ricava che se il numero è espresso con 1 cifra significativa l'errore relativo è compreso tra il 5% e il 50%, mentre con 3 cifre significative l'errore relativo è compreso tra lo 0.05% e lo 0.5%.

Propagazione degli errori

Spesso il risultato di una misura è ottenuto come risultato di una relazione matematica tra grandezze misurate separatamente mediante strumenti. Ad esempio la misura di una tensione può essere ottenuta come prodotto tra due misure indipendenti; una misura di resistenza e una di corrente: V = RI. L'errore sul risultato dipende dagli errori sulle singole misurazioni.

Supponiamo che X e Y siano i valori medi ottenuti su due grandezze con errori pari a ε_X e ε_Y e che tali valori siano utilizzati per calcolare il risultato Z di una grandezza funzione di X e Y. Si pone: Z = f(X,Y)

Si può dimostrare che se:

- la relazione è di somma o sottrazione allora $Z = X \pm Y$ con errore $\varepsilon_Z = \varepsilon_x + \varepsilon_Y$
- la relazione è di prodotto allora $Z = X \cdot Y$ con errore $\varepsilon_Z = \varepsilon_x + \varepsilon_Y$
- la relazione è di quoziente allora $Z = X/Y$ con errore $\varepsilon_Z = \varepsilon_x + \varepsilon_Y$

Si osservi che nei precedenti casi gli errori si sommano.

Esempio.

Due resistenze R_1 = 1000 Ω e R_2 = 500 Ω con tolleranza (errore relativo) al 5% sono collegate in serie. Determinare la resistenza totale.

Risoluzione

La resistenza totale è la somma delle resistenze: $R_T = R_1+R_2$ = 1500 Ω . L'errore assoluto su ciascuna misura vale:

ε_{R1} = 50 Ω e ε_{R2} = 25 Ω. L'errore totale risulta ε_{RT} = 75 Ω. Pertanto si ha:

R_T = (1500 ± 75) Ω con errore relativo percentuale sempre del 5%.

Altri due interessanti casi sono:

- se una misura X con errore ε_X è utilizzata per valutare il prodotto $Z = K \cdot X$ con K costante priva di errore. Si ha:

  $Z = K \cdot X$ con errore $\varepsilon_Z = K \cdot \varepsilon_X$

- se una misura X con errore ε_X è utilizzata per valutare la potenza $Z = X^K$ con K costante priva di errore. Si ha:

  $Z = X^K$ con errore $\varepsilon_Z = K \cdot \varepsilon_X$

Caratteristiche degli strumenti di misura

I parametri caratteristici degli strumenti di misura sono:

- **Linearità**. È la capacità dello strumento di fornire uguali cambiamenti in corrispondenza di uguali variazioni della grandezza da misurare;
- **Portata**. Rappresenta la massima misurazione ammessa;
- **Potere risolutivo**. È la minima variazione della grandezza di ingresso apprezzabile;
- **Stabilità**. È la capacità di mantenere inalterate le proprie caratteristiche di misura;
- **Sensibilità**. È il rapporto tra la variazione dell'indicazione dello strumento e la corrispondente variazione della grandezza di entrata;
- **Classe di precisione**. Indica l'errore relativo percentuale riferito al valore di fondo scala V_{fs}.

$$\varepsilon_{fs}\% = \frac{\varepsilon}{V_{fs}} \cdot 100$$

Le norme CEI dividono gli strumenti in classi: 0.05; 0.01; 0.2; 0.3; 0.5; 1; 1.5; 2.5; 5.

Quelli di classe 0.05 e 0.01 sono strumenti di precisione usati come campione in laboratorio, quelli di classe da 1 a 5 sono gli strumenti portatili o da quadro.

L'errore relativo percentuale $\varepsilon_m\%$ in corrispondente di una generica misura V_m vale:

$$\varepsilon_m\% = \frac{\varepsilon_{fs}\% \cdot V_{fs}}{V_m}$$

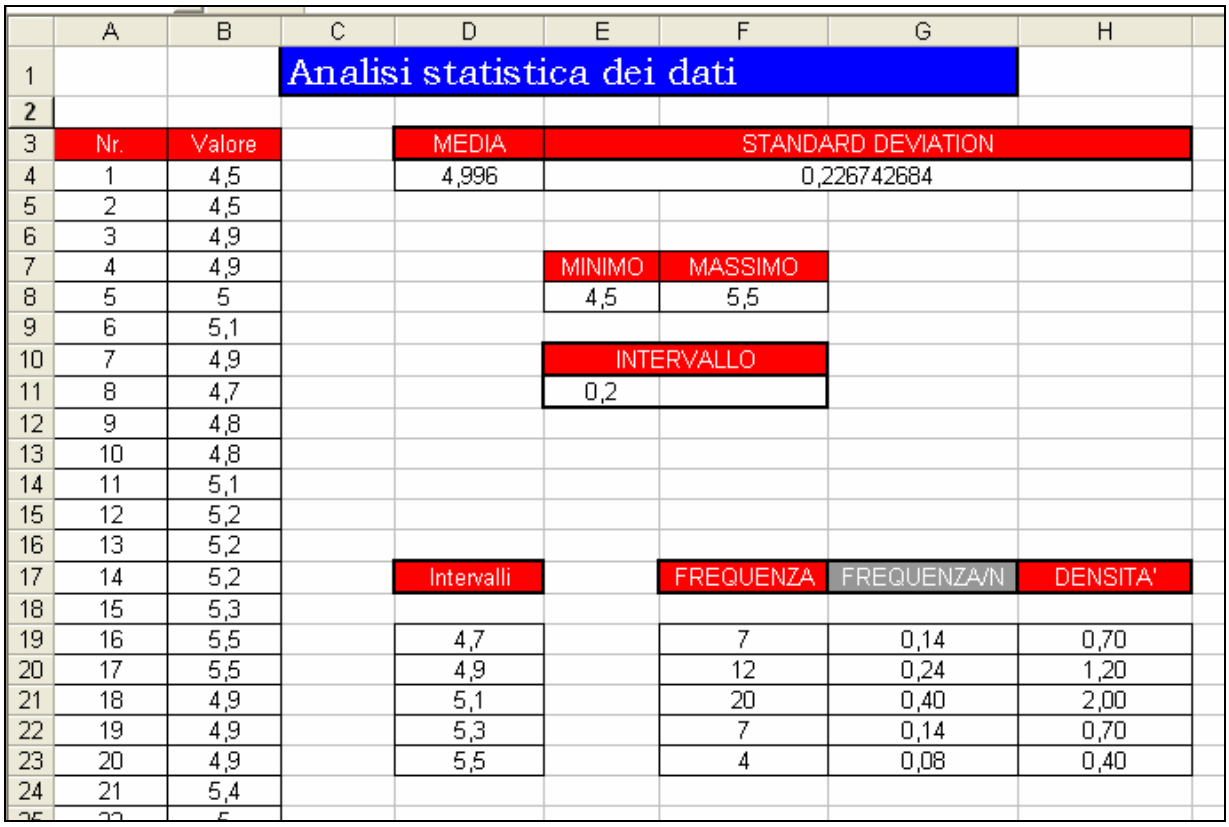
Si comprende che l'errore diminuisce se la misura si effettua vicino al fondo scala.

Misura sperimentale della densità di probabilità

Per ricavare i parametri caratteristici di un insieme di dati sperimentali e determinare il grafico della densità di probabilità si può utilizzare il foglio elettronico e procedere come segue:

- sul foglio elettronico si riportano i valori dei dati sperimentali;
- si calcola la media e la standard deviation;
- si individua l'intervallo entro cui sono distribuiti i dati calcolando il valore minimo e massimo;
- si divide l'intervallo in un certo numero di sotto-intervalli;
- si contano quante misure cadono in ogni sotto-intervalli (frequenza nel sotto-intervallo);
- si divide tale numero per il numero totale delle misure;
- si dividono i numeri ottenuti al passo precedente per l'ampiezza del sotto-intervallo ottenendo così la densità di probabilità relativa a ciascun sotto-intervallo
- si traccia l'istogramma che rappresenta il grafico della densità di probabilità. L'area totale di tale grafico vale, ovviamente, 1 (evento certo).

A titolo di esempio si riporta una parte della struttura di un foglio di lavoro in Excel per l'analisi statistica di dati sperimentali.



	A	B	C	D	E	F	G	H
1			Analisi statistica dei dati					
2								
3	Nr.	Valore		MEDIA	STANDARD DEVIATION			
4	1	4,5		4,996	0,226742684			
5	2	4,5						
6	3	4,9						
7	4	4,9			MINIMO	MASSIMO		
8	5	5			4,5	5,5		
9	6	5,1						
10	7	4,9			INTERVALLO			
11	8	4,7			0,2			
12	9	4,8						
13	10	4,8						
14	11	5,1						
15	12	5,2						
16	13	5,2						
17	14	5,2		Intervalli		FREQUENZA	FREQUENZA/N	DENSITA'
18	15	5,3						
19	16	5,5		4,7		7	0,14	0,70
20	17	5,5		4,9		12	0,24	1,20
21	18	4,9		5,1		20	0,40	2,00
22	19	4,9		5,3		7	0,14	0,70
23	20	4,9		5,5		4	0,08	0,40
24	21	5,4						

Fig. 4. - Analisi statistica di dati sperimentali.

Nella colonna B (non mostrata completamente) da B4 a B53 si sono riportati 50 valori sperimentali.
Nella cella D4 si è valutata la media applicando la formula:

=MEDIA(B4:B53)

Nella cella E4 si calcola la standard deviation applicando la formula:

=DEV.ST(B4:B53)

Nelle celle E8 e F8 si sono scritte le formule per il calcolo del valore minimo e massimo. In formule:

=MIN(B4:B53) **=MAX(B4:B53)**

Nella cella E11 si scrive la formula per il calcolo dell'ampiezza degli intervallini supponendo una divisione in 5 sotto-intervalli:

=(F8-E8)/5

Nelle celle da D19 a D23 si sono inseriti i valori degli intervallini.
Si selezionano le celle di emissione da F19 a F23 e si calcola la frequenza dei dati per ciascun intervallo applicando la formula:

=FREQUENZA(B4:B53;D19:D23)

Ad esempio il numero di dati x sperimentali (frequenza) tra:

$0 < x \leq 4.7$ vale: 7

$4.7 < x \leq 4.9$ vale: 12

……..

$5.3 < x \leq 5.5$ vale: 4

Si osservi che nel primo intervallino cadono tutte le misure tra il valore minimo 4.5 e il valore 4.7 incluso, nel secondo intervallino i valori tra 4.7 e 4.9 incluso e cosi fino all'ultimo intervallino.

Ricordiamo che secondo Bernulli la frequenza è definita come rapporto tra il numero di volte che l'evento si verifica e il numero totale di prove effettuate. In Excel per frequenza si intende il numero di volte che si verifica una misurazione entro un intervallo. Pertanto per calcolare la frequenza secondo Bernulli si devono dividere i valori trovati precedentemente per il numero di prove, nel nostro esempio N = 50

Pertanto, nella cella G19 si scrive la formula:

=F19/50

che si deve trascinare fino alla cella F23.

La densità di probabilità per ciascun intervallino si ottiene dividendo la frequenza di Bernulli di ciascun intervallino per la relativa ampiezza. Nella cella H19 si scrive la formula:

=G19/E11

che si deve trascinare fino alla cella H23.

Attivando la procedura di ***creazione guidata grafico*** si ricava la distribuzione di fig. 5.

È facile verificare che l'area sottesa dal grafico della distribuzione di probabilità vale 1.

Il grafico ottenuto è a gradini poiché è limitato il numero delle misurazioni e dei sotto-intervalli. Se si aumenta il numero delle misurazioni e si rendono sempre più piccoli i sotto-intervalli il grafico diventa sempre più continuo e tende a quello di Gauss.

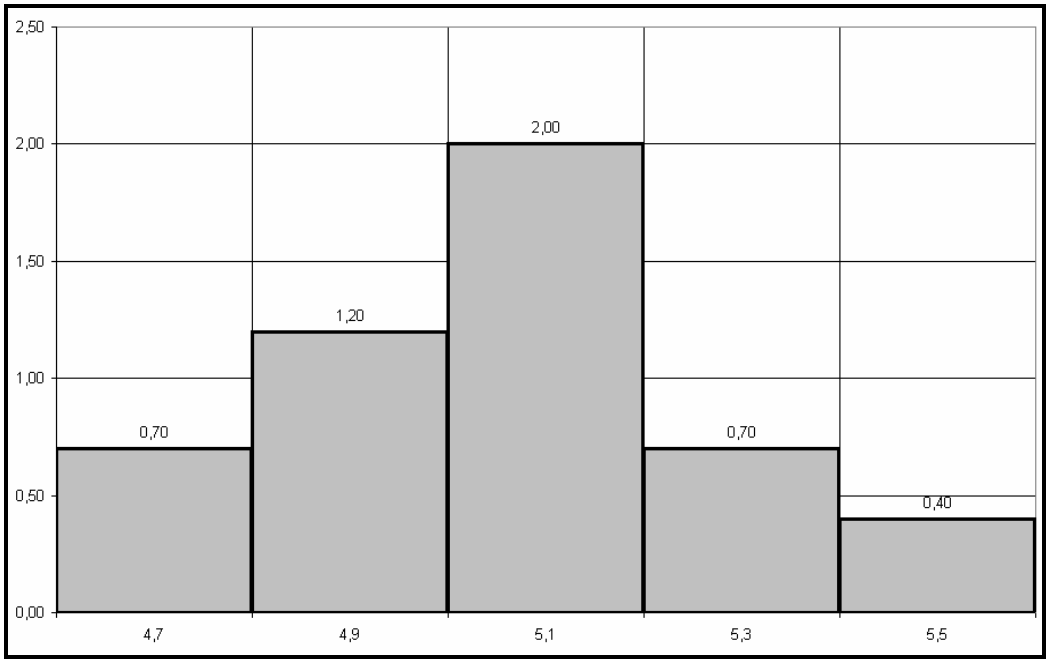


Fig. 5. - Distribuzione dei dati sperimentali.

Distribuzione Esponenziale e Distribuzione di Poisson

Molti fenomeni fisici sono governati da distribuzioni di probabilità diverse da quella di Gauss. Tra queste ricordiamo la distribuzione esponenziale e quella di Poisson.

La **distribuzione esponenziale** trova applicazione in tutti quei fenomeni nei quali si può supporre che la probabilità del verificarsi di un evento è proporzionale alla durata dell'intervallo considerato.

$$p(\Delta t) = \lambda \cdot \Delta t$$

Con λ costante di proporzionalità.

È questo il caso del decadimento radioattivo di una sostanza o lo studio della teoria delle code in cui si hanno partenze e arrivi casuali ad una determinata destinazione come caselli autostradali sportelli bancari, ecc. In questo caso la densità di probabilità è data da:

$$D(t) = \lambda \cdot e^{-\lambda \cdot t}$$

La costante λ ha le dimensioni inverse di un tempo e il suo reciproco $\tau = 1/\lambda$ è la costante di tempo. Per questa distribuzione il valor medio e la deviazione standard coincidono con la costante di tempo.

La **distribuzione di Poisson** è utilizzata, ad esempio, per descrivere la distribuzione dei guasti nei macchinari, nello studio delle trasmissioni dati digitali, nell'analisi del traffico telefonico, ecc. La densità di probabilità è descritta dall'equazione:

$$D(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

Dove λ rappresenta la media e la varianza e $x! = x_1 \cdot x_2 \cdot \ldots\ldots\ldots\ldots \cdot x_N$ è il fattoriale del numero x.

La distribuzione di Poisson è di tipo discreto poiché la varabile x può assumere solo valori interi.

Il test del χ^2 per una distribuzione

Effettuate una serie di misurazioni è fondamentale comprendere quale distribuzione di probabilità teorica si adatta meglio ai dati sperimentali. Un test di verifica è noto come test del **chi-quadro χ^2**. Applichiamo tale metodo alla distribuzione di Gauss.

Supponiamo di aver effettuato **N =40** misure x_1, x_2, ……x_{40} relative ad una certa grandezza fisica. Si valuta il valor medio e la deviazione standard. Supponiamo di avere:

$$x_m = 730.1 \qquad\qquad \sigma = 46.8$$

Per valutare se i dati seguono la distribuzione di Gauss si deve confrontare la distribuzione dei dati osservati O_K sperimentalmente con la distribuzione teorica. Una possibile procedura consiste nel dividere l'intervallo dei dati sperimentali ad esempio in **n = 4** intervalli e calcolare il numero di misurazioni comprese in tali intervalli. Supponiamo di ottenere:

	Intervalli	**Numero di misurazioni O_k**
1	$x < x_m - \sigma$	8
2	$x_m - \sigma < x < x_m$	10
3	$x_m < x < x_m + \sigma$	16
4	$x_m + \sigma < x$	6

Nella seguente tabella si riportano i valori delle misurazioni attese in funzione delle probabilità della distribuzione di Gauss nell'intervallo e quelle osservate sperimentalmente. Assumiamo che tutte le misure siano allocate tra:

- $x_m - 2\sigma < x < x_m + 2\sigma$:con probabilità P_k del 100 %

e che tra:

- $x_m - \sigma < x < x_m + \sigma$:la probabilità P_k sia del 68 %

Si ha:

Intervalli		**Misurazioni Osservate O_k**	**Probabilità P_k nell'intervallo**	**Misurazioni Attese $A_k = N \cdot P_k$**
1	$x < x_m - \sigma$	8	16%	6.4
2	$x_m - \sigma < x < x_m$	10	34%	13.6
3	$x_m < x < x_m + \sigma$	16	34%	13.6
4	$x_m + \sigma < x$	6	16%	6.4

Se i valori sperimentali seguissero perfettamente la distribuzione di Gauss si dovrebbe avere:

$$O_k - A_k = 0$$

Nella pratica per un avere una buona corrispondenza la precedente differenza deve essere la più piccola possibile.

Per valutare se c'è accordo tra la distribuzione osservata e quella attesa si introduce il parametro χ^2 **cosi definito:**

$$\chi^2 = \sum_{K=1}^{n} \frac{(O_k - A_k)^2}{A_K}$$

Se $\chi^2 = 0$ l'accordo è perfetto. Cosa improbabile.

Si dimostra che c'è accordo con la distribuzione teorica se χ^2 è minore o al massimo uguale al numero di intervalli n: $\chi^2 < n$

Se $\chi^2 > n$ non c'è accordo tra la distribuzione sperimentale e quella teorica ipotizzata.

Per l'esempio in esame, applicando la formula precedente, si ricava:

$$\chi^2 = 1.8$$

Poiché tale valore è minore di n = 4 si può ritenere che la distribuzione dei dati sperimentali segue quella di Gauss.

Regressione e Correlazione

Per **analisi di regressione** si intende la ricerca della curva interpolante che meglio approssima la relazione tra le variabili misurate x e y di un sistema presupposto che tra esse vi sia un rapporto di causa-effetto.

L'analisi prevede:

- Raccolta dei dati sperimentali con costruzione di una tabella e di un diagramma di dispersione;
- Ricerca della funzione interpolante y = f(x). La funzione può essere lineare, polinomiale, logaritmica, ecc.;
- Determinazione del coefficiente di correlazione che indica il grado di attendibilità della funzione y = f(x). Cioè il grado di addensamento dei valori sperimentali attorno alla curva di regressione y = f(x).

In fig. 6 si mostra un esempio di tabella e di un grafico di dispersione.

x	y
x1	y1
x2	y2
x3	y3
----------	----------
xi	yi
----------	----------

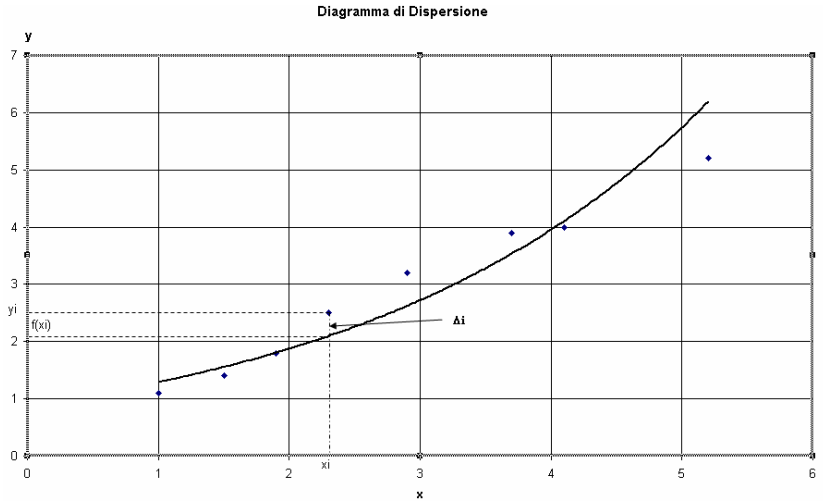


Fig. 6. - Tabella e diagramma di dispersione.

L'analisi della regressione si basa sul **metodo dei minimi quadrati**. Per ogni punto x_i si calcola la deviazione tra il valore yi e il valore $f(x_i)$:

$$\Delta i = yi - f(xi)$$

La curva che meglio approssima la distribuzione dei punti del diagramma di dispersione è quella che rende minima la somma dei quadrati delle deviazioni:

$$\mathbf{min}(\Delta_1^2 + \Delta_2^2 + ...\Delta_n^2)$$

Il metodo prevede l'utilizzo dei quadrati per non tenere conto del segno delle deviazioni.

Sviluppando il metodo dei minimi quadrati si ricavano le formule per il calcolo del coefficiente di correlazione e dei parametri della curva di regressione.

La precedente analisi può essere vantaggiosamente condotta mediante l'uso del foglio elettronico.

Si mostra, in fig. 7, un esempio di analisi di regressione lineare sviluppata in ambiente Excel.

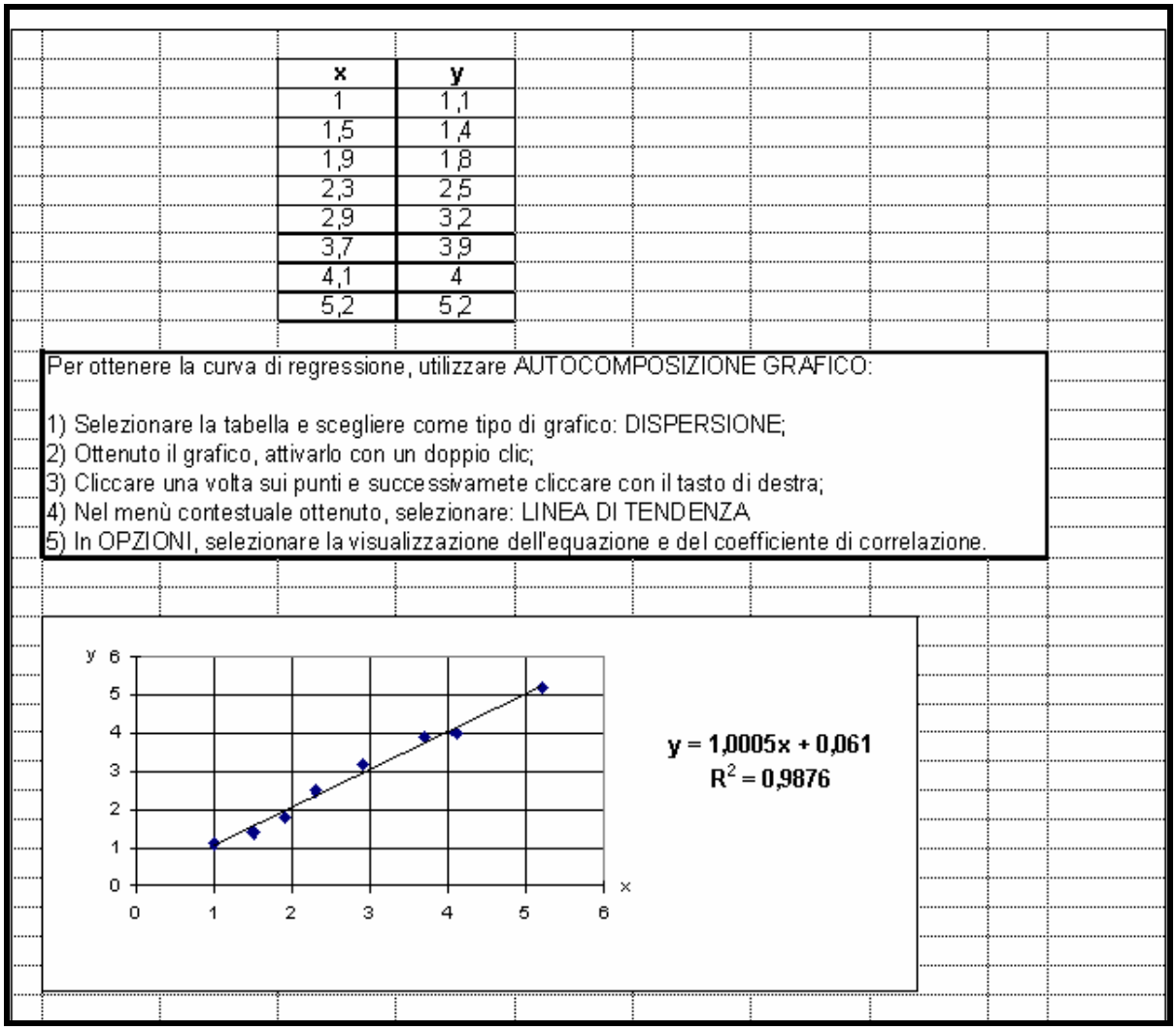


Fig. 7. - Foglio di Excel per l'analisi della regressione e il calcolo del coefficiente di correlazione.

Il coefficiente di correlazione R^2 può assumere valori compresi tra 0 e 1:

- $R^2 = 1$ indica che c'è perfetta correlazione. La funzione di regressione y = f(x) descrive con esattezza assoluta i dati sperimentali;
- $R^2 = 0$ indica che non esiste nessuna correlazione tra funzione y = f(x) e dati sperimentali.

La correlazione si può ritenere buona se $R^2 > 0.5$; ovvero $R > 0.7$ ad indicare che più del 70% della variazione di x può essere espressa dalla curva di regressione y = f(x).