ELEMENTI DI STATISTICA DESCRITTIVA

1. Introduzione

La statistica è la scienza che studia un fenomeno tramite l'organizzazione e l'analisi di dati.

Gli ambiti di utilizzo della statistica sono vari: fisica, medicina, biologia, economia, produzione industriale, demografia, ecc.

Si pensi ad esempio ad una amministrazione comunale che deve impostare il piano regolatore relativo alle nuove costruzioni edilizie sulla base dei flussi di immigrazione-emigrazione. Oppure ad una azienda che deve pianificare la produzione di un oggetto, facendo una analisi di mercato sul tasso di gradimento preventivo di tale oggetto. O ancora lo studio dell'efficacia di un farmaco su un campione di popolazione affetta da una malattia. Tutti questi esempi hanno bisogno di una raccolta di dati, della loro organizzazione e di una analisi dei dati così organizzati, ovvero hanno bisogno di una analisi statistica.

La statistica come la conosciamo oggi nasce nel XIX secolo ad opera principalmente di due scienziati: F.Galton(1822-1911) e K.Pearson(1857-1936) e, più recentemente, Fisher(1890-1962) che si occupò in modo rigoroso della stima statistica e del campionamento. I problemi riguardanti la statistica erano però già studiati nell'antichità: basti pensare ai censimenti effettuati nell'Impero Romano (anche nella Bibbia, parlando della nascita di Cristo, si parla di movimenti di popolazioni che dovevano andare a registrarsi).

Per concludere questa introduzione possiamo dire che la statistica ha una duplice funzione: capire un evento passato tramite la raccolta e l'organizzazione dei dati, cercandone le cause e tentare, in modo più preciso possibile, di prevedere un evento futuro.

2. L'indagine statistica

Vediamo adesso in cosa consiste e a grandi linee una indagine statistica.

Supponiamo di avere una gelateria e di voler pianificare la produzione dei gusti del gelato sulla base delle preferenze espresse dai clienti.

Per prima cosa dobbiamo individuare qual è l'insieme di soggetti che ci daranno le informazioni richieste: in questo caso i clienti della gelateria. Tale insieme di soggetti si chiama **popolazione**. Ogni elemento della popolazione si chiama **unità statistica**.

Se non vogliamo chiedere ad ogni cliente la propria preferenza, ma preferiamo domandarlo solamente ad alcuni clienti, ad esempio quelli abituali, tramite un piccolo questionario, si parlerà allora di campione. Il **campione** è dunque un sottoinsieme della popolazione. In alcuni casi l'utilizzo di un campione è praticamente obbligatorio: si pensi alla intenzione di voto durante una tornata elettorale: è impossibile domandare a tutti i cittadini cosa voteranno, ma verrà chiesto, ad esempio con un sondaggio telefonico, solamente ad un numero ristretto di persone.

Il campionamento di una popolazione è molto importante perché deve rappresentare nel miglior modo possibile tutta la popolazione. Tornando all'esempio della gelateria, supponendo che i clienti siano di tutte le fasce d'età ed equamente distribuiti tra maschi e femmine, un campione non efficace sarà ad esempio quello di soli maschi ultra sessantenni.

Una volta individuata la popolazione o, in alternativa, il campione della popolazione si procede alla **raccolta dei dati**.

La raccolta dei dati può avvenire in molti modi: intervista, questionario scritto, internet, ecc.

Supponiamo che nella nostra gelateria abbiamo un questionario da far riempire ai clienti. La domanda è semplicemente quella di indicare con una crocetta il gusto preferito da una lista. Una volta terminata la raccolta dei dati si effettua lo **spoglio**, ovvero la lettura dei dati e la loro trascrizione su una tabella. Di seguito un esempio:

Gusto	Frequenza assoluta	Frequenza relativa	Frequenza relativa in %
Cioccolato	8	0,175	17,50 %
Panna	6	0,130	13,00 %
Vaniglia	4	0,087	8,70 %
Crema	7	0,152	15,20 %
Stracciatella	5	0,109	10,90 %
Fragola	4	0,087	8,70 %
Pistacchio	6	0,130	13,00 %
Cassata	4	0,087	8,70 %
Cocco	2	0,043	4,30 %
Totale	**46**	**1**	**100 %**

Nella seconda colonna si mette la **frequenza assoluta**, cioè il numero di unità statistiche per ciascun gusto. Talvolta è utile indicare anche la **frequenza relativa**, nella terza colonna, che è il rapporto tra la frequenza assoluta e il totale di unità statistiche. La somma delle frequenze relative è sempre 1.

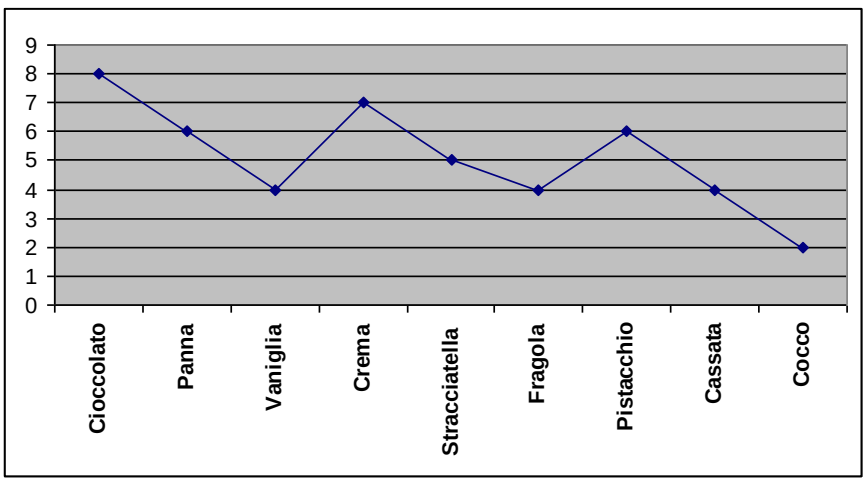
La frequenza relativa può essere rapportata a 100, ogni dato è dunque descritto da una percentuale. Per ottenere la percentuale è sufficiente moltiplicare la frequenza relativa per 100 (quarta colonna).

3. La rappresentazione grafica

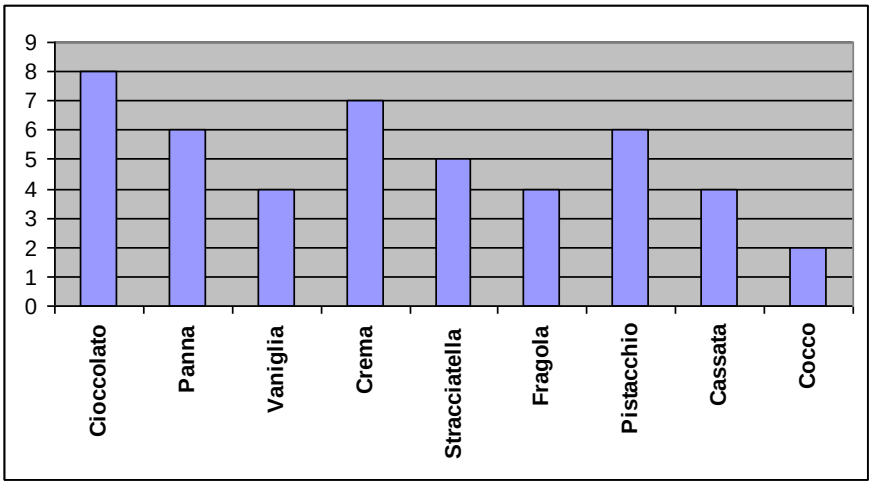
La rappresentazione grafica è un metodo alternativo rispetto a quello tabellare di rappresentare i dati raccolti. Il vantaggio della rappresentazione grafica è la facilità di lettura e comprensione.

I principali tipi di diagrammi sono: **grafici cartesiani**, **grafici a barre**, **areogrammi**. (Le denominazioni possono cambiare).

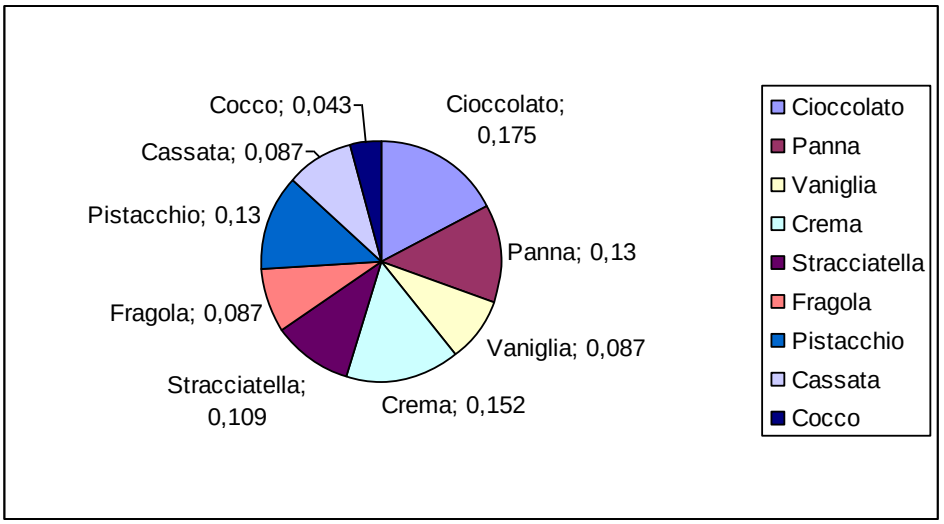
Nei **grafici cartesiani** si riporta sull'asse delle ascisse le tipologie e sull'asse delle ordinate le frequenze.

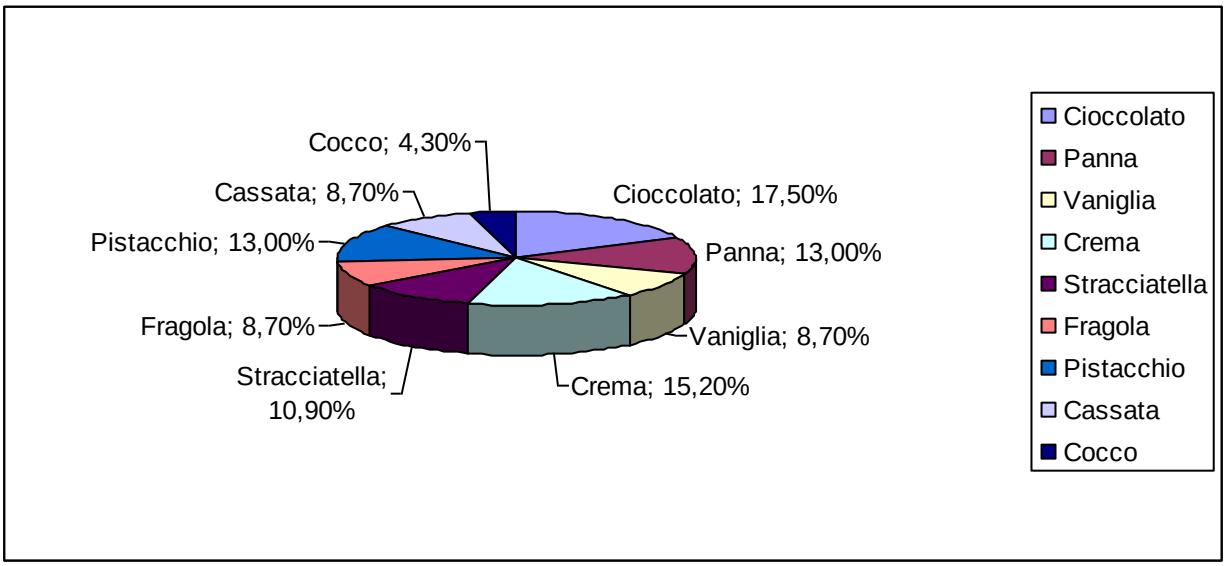


Nei **grafici a barre**, che possono essere verticali o orizzontali, si fissano dei rettangoli di uguale base e altezza proporzionale alle frequenze.



Negli **areogrammi**, alcuni chiamati grafici a torta, si rappresentano i vari spicchi di area proporzionali alle frequenze.





Generalmente nei primi due tipi di grafici si riportano le frequenze assolute, nei grafici a torta le frequenze relative.

La tipologia di grafico da scegliere dipende dai tipi di dato. Ad esempio se le tipologie sono poche vanno bene i grafici a barre, se sono molte sono preferibili i grafici cartesiani.

3. Gli indici di posizione

Con gli indici di posizione si inizia quella possiamo definire l'analisi dei dati raccolti.

MEDIA

Definizione: si chiama **media aritmetica semplice** di n numeri $x_1, x_2, ..., x_n$ il rapporto fra la loro somma ed n:

$$M = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Esempio: nella nostra rilevazione la media è data da $M = \frac{8+6+4+7+5+4+6+4+2}{9} = 5{,}11$

Possiamo allora concludere che in media i gusti hanno avuto una preferenza da 5,11 unità statistiche.

Per completezza scriviamo le definizioni di altre medie.

Definizione: si chiama **media ponderata** di n numeri $x_1, x_2, ..., x_n$, ciascuno associato ad un altro numero chiamato peso, la somma dei prodotti di ciascun numero con il proprio peso diviso la somma dei pesi:

$$M = \frac{p_1 \cdot x_1 + p_2 \cdot x_2 + ... + p_n \cdot x_n}{p_1 + p_2 + ... + p_n} = \frac{\sum_{i=1}^{n} p_i \cdot x_i}{\sum_{i=1}^{n} p_i}$$

Definizione: si chiama **media geometrica semplice** di n numeri $x_1, x_2, ..., x_n$ la radice n-esima del loro prodotto:

$$M_G = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Definizione: si chiama **media armonica semplice** di n numeri $x_1, x_2, ..., x_n$ il reciproco della media aritmetica dei reciproci dei numeri.:

$$M_A = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

MODA

Definizione: si chiama **moda o valore modale** di una distribuzione di frequenza il termine, se esiste, cui corrisponde la massima frequenza.

Nel nostro esempio è il 4, perché si ripete 3 volte, mentre gli altri valori si ripetono non più di 2 volte.

MEDIANA

Definizione: si chiama **mediana** di una distribuzione di frequenza il termine che, dopo aver ordinato i dati in ordine crescente o decrescente, occupa il posto centrale.

Nel nostro esempio dobbiamo prima ordinare i dati:

2-4-4-4-**5**-6-6-7-8

La mediana è dunque il 5 perché occupa il posto centrale

Nel caso in cui i dati siano pari, si prende la media aritmetica dei due termini centrali, anche se non appartiene alla distribuzione.

Esempio:

2-3-5-5-**6**-**8**-9-12-14-15

i termini centrali sono 6 ed 8, quindi per mediana si prende il valore 7.

4. Gli indici di variabilità

Gli indici di posizione hanno il vantaggio di riassumere in un unico numero tutta la distribuzione che si deve studiare, ma hanno lo svantaggio di non tenere conto di come sono distribuiti i numeri intorno ad essi. Ad esempio la serie: 6,6,6,6,6,6,6 ha media 6 e tutti i valori coincidono con la media stessa. Quindi la variazione dei valori intorno alla media è nulla.

Anche la serie 1,3,5,6,7,9,11 ha media 6, ma quasi tutti i valori della serie si discostano sensibilmente da tale valore. Quindi la variazione dei valori intorno alla media non è nulla.

Per descrivere queste variazioni dobbiamo introdurre altri numeri detti “indici di variabilità”.

CAMPO DI VARIAZIONE

Definizione: si chiama **campo di variazione** di una distribuzione di frequenza la differenza fra il numero maggiore e il minore.

Nel nostro esempio il numero maggiore è 8 e il minore 2, quindi il campo di variazione è 6.

SCARTO ASSOLUTO DALLA MEDIA

Definizione: si chiama **scarto assoluto dalla media** S_i di un valore x_i di una distribuzione di frequenza $x_1, x_2, ..., x_n$ il valore assoluto della differenza di x_i e la media aritmetica M: $|x_i - M|$

Nel nostro esempio il gusto cocco, che ha una frequenza 2, ha uno scarto assoluto dalla media (5,11) di:

$$|x_i - M| = |2 - 5{,}11| = |-3{,}11| = 3{,}11$$

SCARTO SEMPLICE MEDIO

Definizione: si chiama **scarto semplice medio** S di una distribuzione di frequenza $x_1, x_2, ..., x_n$ la media aritmetica degli scarti assoluti dalla media:

$$S = \frac{|x_1 - M| + |x_2 - M| + ... + |x_n - M|}{n} = \frac{\sum_{i=1}^{n} |x_i - M|}{n}$$

Nel nostro esempio lo scarto semplice medio è dato da:

$$S = \frac{|8-5{,}11| + |6-5{,}11| + |4-5{,}11| + |7-5{,}11| + |5-5{,}11| + |4-5{,}11| + |6-5{,}11| + |4-5{,}11| + |2-5{,}11|}{9} = 1{,}46$$

Questo vuol dire che mediamente i valori si discostano dalla media di 1,46. Più S è grande, più i valori si discostano dalla media.

Osservazione: gli scarti devono essere presi in valore assoluto, altrimenti si dimostra facilmente che S varrebbe sempre 0.

Più frequentemente, rispetto agli indicatori precedenti, vengono usati quelli descritti di seguito.

SCARTO QUADRATICO

Definizione: si chiama **scarto quadratico** di un valore x_i di una distribuzione di frequenza $x_1, x_2, ..., x_n$ il quadrato della differenza di x_i e la media aritmetica M: $(x_i - M)^2$

Nel nostro esempio il gusto cocco, che ha una frequenza 2, ha uno scarto assoluto dalla media (5,11) di: $(x_i - M) = (2-5{,}11)^2 = (-3{,}11)^2 = 9{,}67$

VARIANZA

Definizione: si chiama **varianza** σ^2 di una distribuzione di frequenza $x_1, x_2, ..., x_n$ la media aritmetica degli scarti quadratici:

$$\sigma^2 = \frac{(x_1 - M)^2 + (x_2 - M)^2 + ... + (x_n - M)^2}{n} = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n}$$

Nel nostro esempio la varianza è data da:

$$\sigma^2 = \frac{(8-5{,}11)^2 + (6-5{,}11)^2 + (4-5{,}11)^2 + (7-5{,}11)^2 + (5-5{,}11)^2 + (4-5{,}11)^2 + (6-5{,}11)^2 + (4-5{,}11)^2 + (2-5{,}11)^2}{9} = 2{,}99$$

SCARTO QUADRATICO MEDIO (o DEVIAZIONE STANDARD)

Definizione: si chiama **scarto quadratico medio (o deviazione standard)** σ di una distribuzione di frequenza $x_1, x_2, ..., x_n$ la radice quadrata della varianza:

$$\sigma = \sqrt{\frac{(x_1 - M)^2 + (x_2 - M)^2 + ... + (x_n - M)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}}$$

Nel nostro esempio lo scarto quadratico medio è dato da:

$$\sigma = \sqrt{\frac{(8-5{,}11)^2 + (6-5{,}11)^2 + (4-5{,}11)^2 + (7-5{,}11)^2 + (5-5{,}11)^2 + (4-5{,}11)^2 + (6-5{,}11)^2 + (4-5{,}11)^2 + (2-5{,}11)^2}{9}} = \sqrt{2{,}99} = 1{,}73$$

6. Distribuzioni di probabilità

Supponiamo di prendere un mazzo di 40 carte e di fare un gioco. Si pesca una carta e se esce cuori vinco 2 euro, se esce quadri vinco 1 euro, se esce fiori o picche perdo 3 euro.
Si può riassumere questo esperimento nella seguente tabella (dove le vincite sono espresse con numeri positivi e le perdite con numeri negativi):

E	**X**	**p**
Carta di cuori	2	¼
Carta di quadri	1	¼
Carta di fiori o picche	-3	½

E indica l'evento:

evento E_1: "viene estratta una carta di cuori"

evento E_2: "viene estratta una carta di quadri"

evento E_3: "viene estratta una carta di fiori o picche"

X è la **variabile aleatoria** che, nel nostro esempio, può assumere i valori dell'insieme {2,1,-3}:

$x_1=2$

$x_2=1$

$x_3=-3$

p è la probabilità che si verifichi l'evento:

$p(E_1)=¼$

$p(E_2)=¼$

$p(E_3)=½$

Osservazione:

I singoli eventi costituiscono una partizione sull'insieme di tutti gli eventi, cioè:

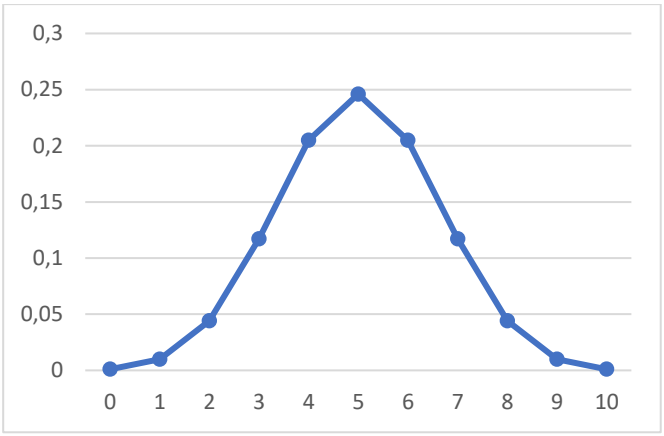
- Nessun evento ha probabilità nulla, ovvero almeno uno dei tre eventi deve verificarsi
- Sono **incompatibili**, ovvero a due a due disgiunti
- Sono **complementari**, ovvero la loro unione dà tutto l'insieme degli eventi. Questo significa che la somma delle probabilità è 1.

La **variabile aleatoria** (o variabile casuale) è dunque una quantità variabile che può assumere certi valori al verificarsi di eventi che soddisfano le condizioni esposte nell'osservazione.
Nel nostro esempio l'insieme dei valori che può assumere la variabile aleatoria è un sottoinsieme degli Interi: si parla dunque di **variabile aleatoria discreta**. Nel caso in cui l'insieme dei valori fosse un sottoinsieme dei Reali si parla **variabile aleatoria continua**: ad esempio le altezze di un gruppo di individui, il peso di barre di ferro (supponendo di avere strumenti di precisione infinita).

Facciamo un altro esempio. Supponiamo di lanciare 10 volte una moneta e registrare quante volte viene Testa. L'insieme dei valori che può assumere la variabile aleatoria è: {0,1,2,3,4,5,6,7,8,9,10}. Calcolando la probabilità di ciascun evento si ha la seguente tabella:

E	X	p
Esce 0 volta Testa	0	0,001
Esce 1 volta Testa	1	0,010
Esce 2 volta Testa	2	0,044
Esce 3 volta Testa	3	0,117
Esce 4 volta Testa	4	0,205
Esce 5 volta Testa	5	0,246
Esce 6 volta Testa	6	0,205
Esce 7 volta Testa	7	0,117
Esce 8 volta Testa	8	0,044
Esce 9 volta Testa	9	0,010
Esce 10 volta Testa	10	0,001

Rappresentando i valori in un grafico cartesiano nel quale l'asse delle ascisse è rappresentato da **X** e l'asse delle ordinate da **p**, si ha:

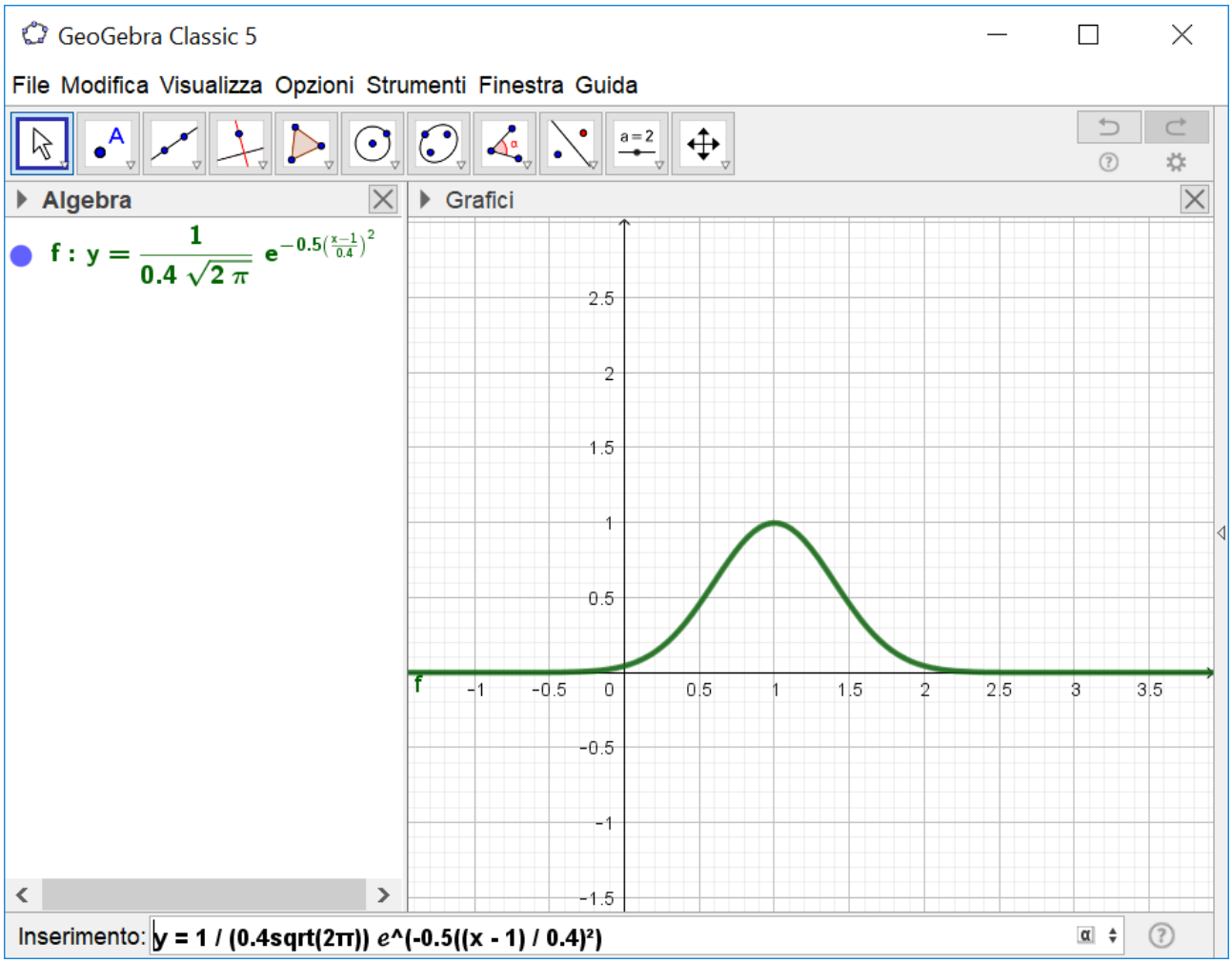


Si nota che la funzione ha una forma "a campana". Una **distribuzione** di questo tipo si dice normale o **gaussiana**, dal nome del "princeps mathematicorum" **Carl Friederich Gauss** (1777-1855).
La distribuzione gaussiana è talmente famosa, che nei 10 marchi tedeschi (prima dell'avvento dell'euro), con l'immagine di Gauss era rappresentata proprio questa funzione:

L'espressione analitica della gaussiana è: $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

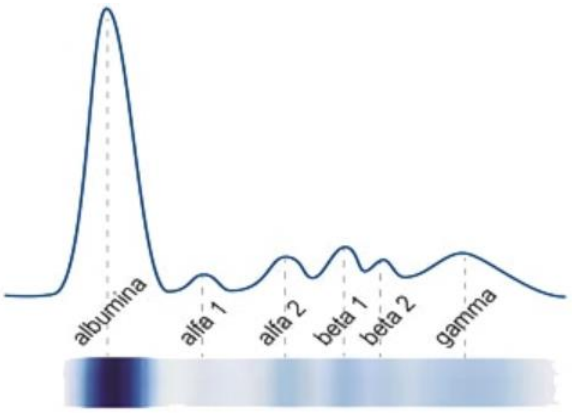
Dove μ è la media e σ lo scarto quadratico medio.

Esempio in Geogebra di una gaussiana:

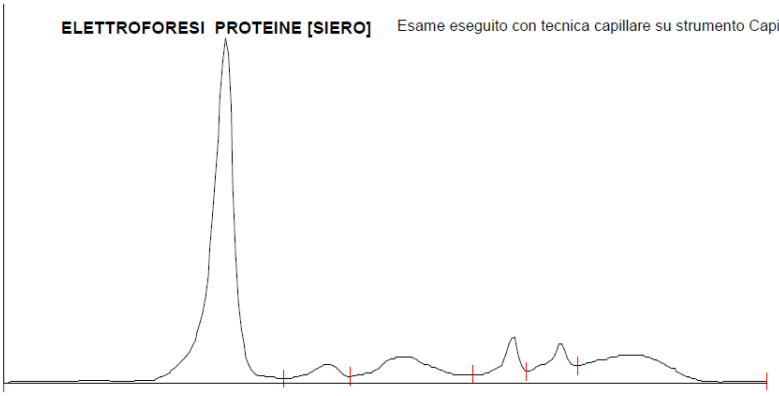


Esempi che danno luogo a una distribuzione gaussiana (o simil-gaussiana):

1. Supponiamo di considerare l'altezza degli italiani maschi. Analizziamo un campione di 1.000 soggetti. Probabilmente otterremmo una curva a campana, centrata attorno a una media, del tipo 174 cm di media con una "deviazione standard" di circa 20 cm, cioè il 95% dei soggetti analizzati sarebbe compreso fra 154 cm e 194 cm.
2. Elettroforesi delle sieroproteine



Protidogramma elettroforetico



Frazioni	%		Int.rif. %	g/dl		Int.rif g/dl
Albumina	58.10		[55.80 - 66.10]	4.4		[3.6 - 5.4]
Alfa 1 Globuline	4.00		[2.90 - 4.90]	0.30		[0.18 - 0.40]
Alfa 2 globuline	10.90		[7.10 - 11.80]	0.80		[0.45 - 0.96]
Beta 1 globuline	6.1		[4.7 - 7.2]	0.50		[0.34 - 0.52]
Beta 2 globuline	6.6	**	[3.2 - 6.5]	0.50	**	[0.23 - 0.47]
Gamma globuline	14.30		[11.10 - 18.80]	1.10		[0.71 - 1.54]

3. La funzione d'onda descritta da Max Born in meccanica quantistica:

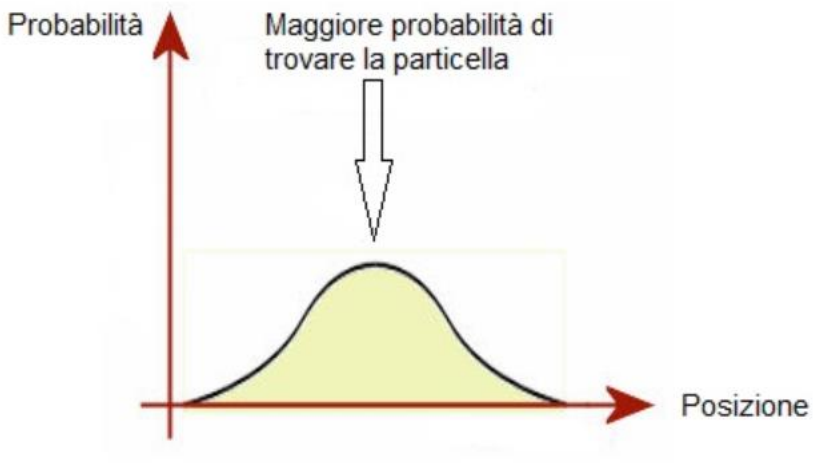


Fig. 1 Funzione d'onda che esprime la probabilità di trovare l'elettrone in un volume infinitesimo di spazio

Altre distribuzioni di probabilità sono:

- Distribuzione binomiale (o di Bernoulli): $P_r = \binom{n}{r} p^r q^{n-r}$
- Distribuzione di Poisson: $P_n(t) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$
- Distribuzione esponenziale $f(t) = \lambda e^{-\lambda t}$

6. Fonti - Risorse

Sitografia

Di seguito una lista con approfondimenti sulla statistica descrittiva e indagini statistiche:

http://www.dmmm.uniroma1.it/~angelo.gilio/stdinfo/esami-Civile-Trasporti/statisticadescrittiva.pdf

http://web.math.unifi.it/users/francini/statisticadescrittiva.pdf

http://www.slideshare.net/ESmargiassi/appunti-statistica-descrittiva-1

http://www.istat.it/it/

http://liceocuneo.it/de-bernardi/wp-content/uploads/sites/13/DISTRIBUZIONI-DI-PROBABILITA.pdf

https://www.albanesi.it/raziologia/curva-di-gauss.htm

https://eterodossia.com/linterpretazione-della-meccanica-quantistica

Bibliografia

Bergamini-Trifone-Barozzi, Matematica verde Vol1 – Zanichelli

Re Fraschini-Grazzi-Spezia, Matematica per l'economia tomo E Statistica e Probabilità – Atlas

Lindley, Introduction to probability and statistics – Cambridge University Press