Consiglio Nazionale delle Ricerche
Istituto di Ricerche sulla Popolazione e le Politiche Sociali

QUANDO LA SCIENZA È TRASVERSALE

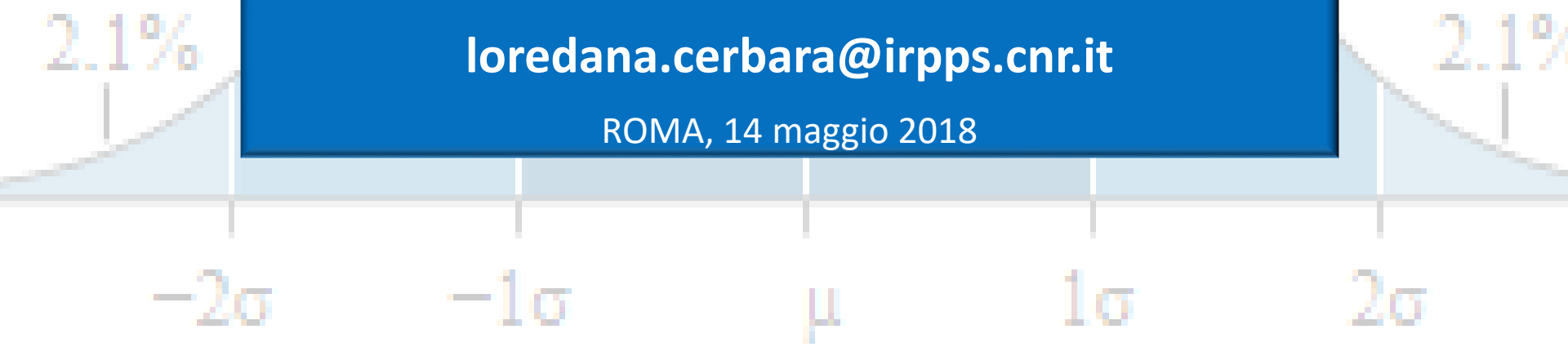
Problem solving e interdisciplinarietà delle competenze matematico - scientifiche a scuola

La statistica senza barriere: uso quotidiano (in)consapevole

Loredana Cerbara

loredana.cerbara@irpps.cnr.it

ROMA, 14 maggio 2018



La statistica non fa per me

Non è raro che si pronunci questa fatidica frase. Eppure tutti, prima o poi, abbiamo a che fare con un dato statistico, usiamo un termine tecnico, comprendiamo un valore probabilistico.

Quando?

Ogni volta che scommettiamo, ogni volta che leggiamo sui quotidiani i risultati elettorali, ogni volta che ci informiamo sulle previsioni del tempo, ogni volta che andiamo dal dottore, ogni volta che …….

La maggior parte delle volte capiamo quanto basta per farci usare il dato senza però sapere come è stato prodotto.

La statistica non fa per me

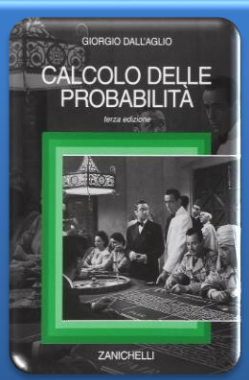
Gottfried Achenwall (Elbląg, 20 ottobre 1719 – Gottinga, 1º maggio 1772) è stato un giurista, storico e filosofo tedesco dell'università di Gottinga

Suo il merito di aver coniato la parola **statistica** ovvero la disciplina che ha lo scopo di descrivere le cose notevoli dello Stato, e di essersi occupato delle sue fondamenta specialmente con il trattato «La costituzione degli stati europei nelle sue linee fondamentali» del 1752 che ebbe ben sette edizioni.

La statistica non fa per me

Ma ci sono tracce di calcoli statistici fin dai primi documenti che sono pervenuti attraverso i millenni. Si trovano tracce di rilevazioni statistiche in Egitto fin dal 3000 a.C., in Mesopotamia, in Cina, presso il popolo ebraico, nell'antica Roma. Anche nel Medioevo e fino ai giorni nostri.

Appena le società si dotarono una organizzazione, i governanti sentirono la necessità di conoscere l'ammontare della popolazione, e in particolare degli uomini e dei soldati, e quello dei beni disponibili → **STATISTICA DESCRITTIVA**



Inoltre da sempre si è sentita la necessità di prevedere i fenomeni e di rifersi a modelli di riferimento utili sia per comprendere l'andamento dei fenomeni, sia a prevederne l'evoluzione

→ **CALCOLO DELLE PROBABILITA' E INFERENZA**

Statistica descrittiva

Esercizio

Un gruppo di 5 studenti è stato selezionato per un test. Il risultato è il seguente:

Studente 1 voto 9
Studente 2 voto 10
Studente 3 voto 8
Studente 4 voto 10
Studente 5 voto 9

Il voto medio ottenuto da questi studenti è: $\frac{9+10+8+10+9}{5} = 12$

Cosa si può concludere?

Statistica descrittiva

1. Alcuni non sono attenti a ciò che sto dicendo

Ho commesso un errore di calcolo evidente!

2. Il valore ottenuto non può essere il valore medio perché è superiore al massimo valore della distribuzione su cui è calcolato. Quindi, anche senza fare calcoli si può intuire che il risultato non è corretto.

3. Il valore centrale di una distribuzione, di cui la media aritmetica è il maggiore rappresentante, è uno dei cardini della statistica descrittiva che serve a fornire visioni dei dati che riescono a superare i limiti umani.

Statistica descrittiva

Infatti, se dico che 5 studenti hanno eseguito un test ed hanno preso una media di 9,2 su 10, è immediato pensare che gli studenti sono stati bravi perché il valore centrale è tanto alto

Se poi aggiungo che lo studente meno bravo ha avuto 8, sembra di non aver bisogno di altre informazioni. Ho aggiunto un'informazione sul range di variabilità dei dati.

Dunque con 2 valori ho sintetizzato una distribuzione di 5 valori. **Il vantaggio di queste sintesi è tanto maggiore quanto più numeroso è il collettivo su cui calcolo la distribuzione.**

Inferenza

Supponiamo che i 5 studenti di prima fossero selezionati da una intera classe. Si può dire che essi rappresentano tutta la classe? Cioè: posso affermare che la media di 9,2 può essere assegnata a tutta la classe anziché ai soli 5 studenti?

La risposta è sì, ma sotto alcune imprescindibili condizioni.

Prima di tutto devo capire se si tratta o no di un campione probabilistico. Infatti solo in questo caso sono autorizzata a riportare i risultati calcolati sul campione all'universo da cui esso deriva.

Statistica descrittiva

Esercizio parte 2

Se la classe era composta da 15 studenti con i seguenti risultati:

Studente 1 voto 9	Studente 6 voto 6	Studente 11 voto 9
Studente 2 voto 10	Studente 7 voto 4	Studente 12 voto 10
Studente 3 voto 8	Studente 8 voto 5	Studente 13 voto 6
Studente 4 voto 10	Studente 9 voto 7	Studente 14 voto 7
Studente 5 voto 9	Studente10 voto 2	Studente 15 voto 9

Il voto medio ottenuto da questi studenti è:

$$\frac{9+10+8+10+9+6+4+5+7+2+9+10+6+7+9}{15} = 7{,}4$$

Cosa si può concludere considerando che sono stati presi solo i primi 5 studenti?

Inferenza

1. Il campione è distorto, cioè ci riporta una situazione più rosea di quanto sia in realtà

2. Avrei potuto scegliere una diversa combinazione di studenti. Avevo a disposizione $\binom{15}{5} = \frac{15!}{5!\cdot 10!} = 3003$ possibili gruppi di 5 studenti presi da una classe di 15. Da ciascuno di questi campioni avrei ottenuto una media e la media di tutte le 3003 medie avrebbe avuto valore 7,4.

Qui risiede la fiducia che ho in un campione!

Ciò può essere avvenuto per pura casualità oppure per un errore sistematico, cioè ho selezionato male gli studenti ed essi non sono 'rappresentativi' dell'intera classe.

Con una procedura casuale avrei avuto una probabilità su 3003 di estrarre questo campione.

Inferenza

Un campione è di tipo probabilistico che per ottenerlo ho usato una **estrazione casuale** delle unità. Esistono molti tipi di campione probabilistico, anche molto complessi, ma almeno in una delle fasi della sua estrazione deve essere intervenuta la probabilità.

Corollario: se non c'è alcun momento dell'estrazione del campione in cui si segue un procedimento casuale il campione non è probabilistico e nulla si può dire sul suo rapporto con l'universo da cui è estratto

Conseguenza: molti dei campioni a cui siamo abituati NON sono probabilistici (indagini telefoniche, indagini postali, indagini senza struttura di campionamento….)

Inferenza

Vantaggi di un campione probabilistico: si può calcolare l'errore di campionamento, cioè posso associare un livello di fiducia alle stime che calcolo sul campione e misurare il rischio che corro nel riportare quelle stime all'universo di partenza.

Svantaggi di un campione probabilistico: non è semplice rispettare un disegno campionario, comporta costi di esecuzione e la disponibilità di una lista di partenza da cui estrarre le unità da inserire nel campione.

In sintesi: c'è un rischio di sbagliare, ma posso calcolare la probabilità di errore a fronte di uno sforzo maggiore per ottenere il campione

Inferenza

Un caso reale: le elezioni politiche di marzo 2018

Prima delle elezioni:

Dati effettivi:

CentroDestra 221

CentroSinistra 112

M5S 221

LeU 14

Tre giorni prima la situazione rilevata attraverso le intenzioni di voto era molto diversa dal risultato finale. Tuttavia già si intravvedeva la difficoltà che poi abbiamo potuto constatare.

Generalmente si tratta di sondaggi telefonici su 1000 casi

Inferenza

Un caso reale: le elezioni politiche di marzo 2018

Primi exit poll e risultati finali

	LA7	RAI	MEDIASET
MOVIMENTO 5 STELLE	28,8 - 30,8	29,5 - 32,5	29,0 - 33,0
FORZA ITALIA	13,5 - 15,5	12,5 - 15,5	12,0 - 16,0
LEGA	12,3 - 14,3	12,5 - 15,5	12,0 - 16,0
FRATELLI D'ITALIA	4,4 - 5,4	3,5 - 5,5	4,0 - 6,0
NOI CON L'ITALIA - UDC	1,8 - 2,4	1,0 - 3,0	0,5 - 2,5
PARTITO DEMOCRATICO	21,0 - 23,0	20,0 - 23,0	17,5 - 21,5
+EUROPA	2,8 - 3,4	2,0 - 4,0	2,0 - 4,0
CIVICA POPOLARE LORENZIN	0,4 - 1,0	0,0 - 2,0	0,4 - 1,0
ITALIA EUROPA INSIEME	0,5 - 1,1	0,0 - 2,0	0,5 - 1,1
LIBERI E UGUALI	5,2 - 6,2	3,0 - 5,0	3,0 - 5,0

M5S 32,4

FI 14,0

Lega 17,6

PD 18,8

FdI 3,3

LeU 3,4

Generalmente sono dati con una probabilità di errore di 1% o 2%

Inferenza

Un caso reale: le elezioni politiche di marzo 2018

Non si tratta solo di errori di campionamento, ma anche di impostazione delle indagini. La probabilità ha un ruolo marginale ma è obbligatorio esplicitarla. Eppure pochi ci fanno caso e tutti sanno che questi dati vanno considerati con cautela. In effetti tutti sono consapevoli che si fa un atto di fiducia nel considerarli e che ci si può sbagliare.

Ma sempre di poco.

Altra cosa sono i sondaggi senza alcun fondamento statistico di cui sono piene le pagine web e i giornali

La distribuzione normale

Come la media aritmetica è il caposaldo della statistica descrittiva, così la distribuzione normale lo è della probabilità e dell'inferenza.

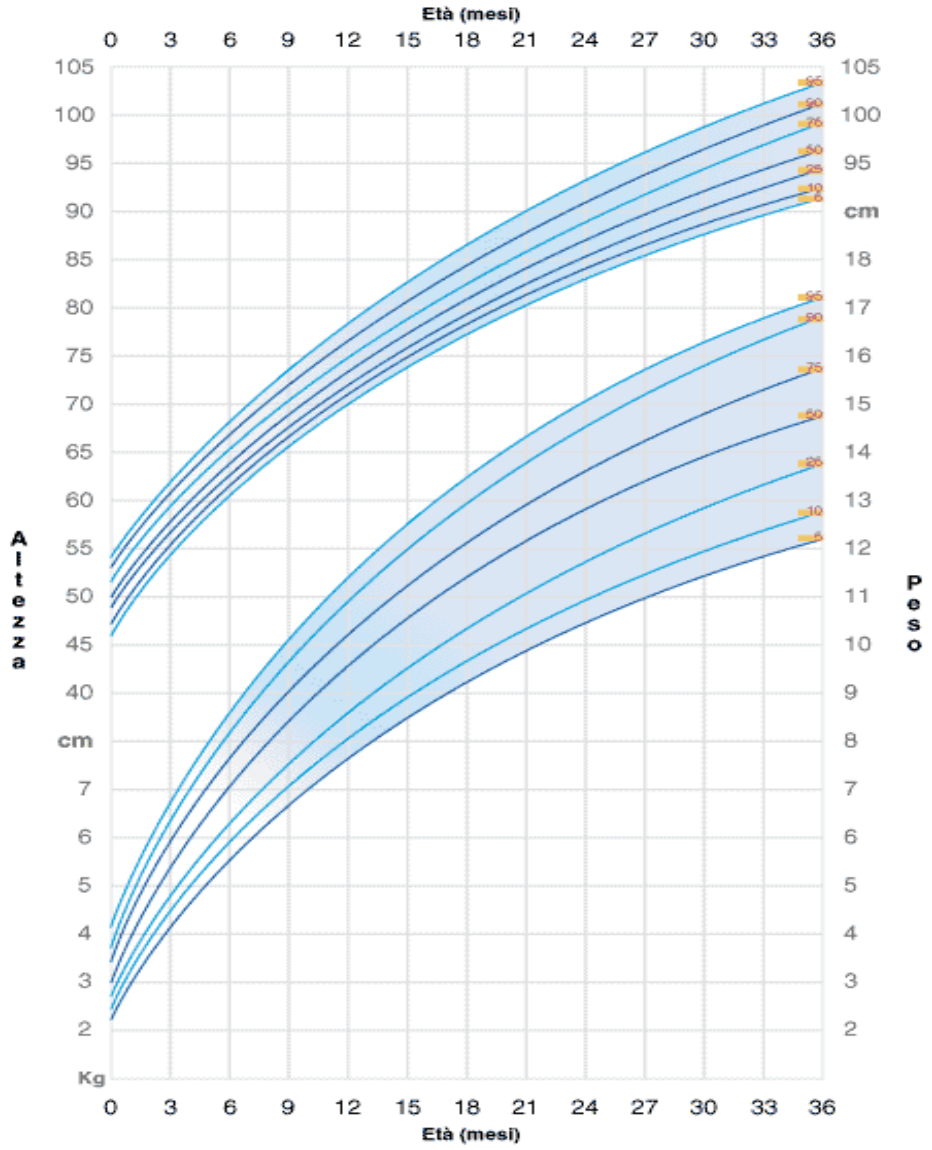
A lei ci riferiamo in molteplici situazioni, tanto che, quella che era stata inventata dal matematico Gauss e che per questo era chiamata Gaussiana, è diventata '**normale**'

È anche detta '**distribuzione degli errori accidentali**' perché è la distribuzione di probabilità che mette ordine quando gli errori sono così tanti da non poterne tener conto.

La somma di tante casualità incontrollate è normale

La distribuzione normale

È la distribuzione a cui si fa riferimento quando si parla di misure biometriche. Ogni mamma sa che il proprio bambino è ‘normale’ se si può inquadrare entro limiti stabiliti da una curva di probabilità e il pediatra indicherà un percentile, compreso tra 0 e 100 per stimare il livello di normalità del bambino.



La distribuzione normale

Un esercizio sulla normale

In una scuola media nella città di Sapri, è stato fatto un esperimento: spiegare ai ragazzi di 12 anni cosa è la distribuzione di probabilità normale.

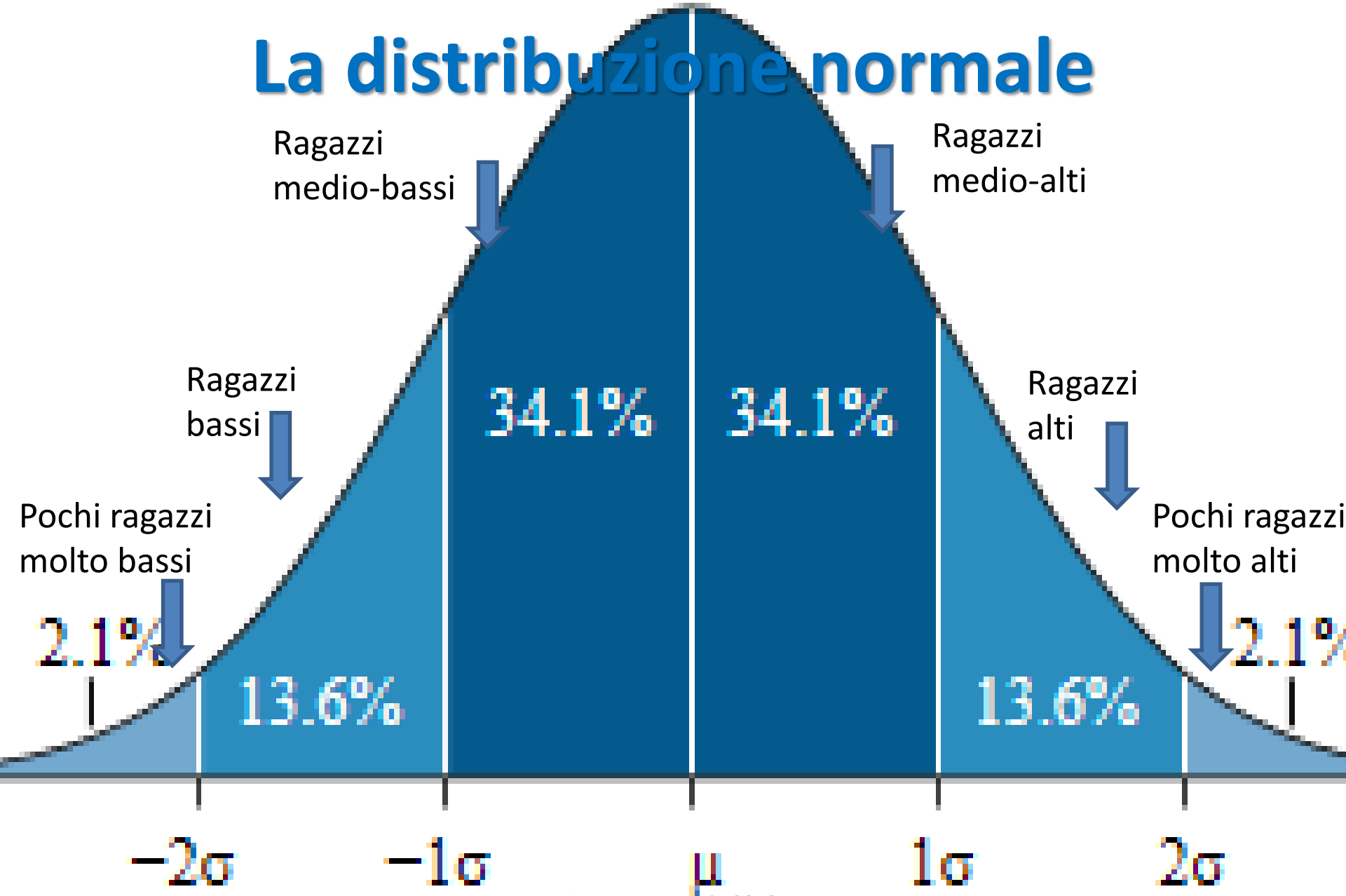
I ragazzi sono stati divisi in più squadre, a ciascuna è stato affidato uno strumento di misura della statura. Sono state rilevate le stature di maschi e femmine ed è stata compilata una tabella con tutte le misure.

La distribuzione normale

Un esercizio sulla normale

Dopo aver osservato la distribuzione, si è deciso di formare 6 livelli di statura e gli studenti sono stati raggruppati secondo questi livelli. I più bassi a sinistra, poi via via sempre più alti. Gli altissimi a destra. Lo spazio era grande e i ragazzi hanno potuto disporsi su 6 file, formando proprio la distribuzione normale.

La distribuzione normale



La distribuzione normale

Lo stesso esperimento, ma con un questionario online autocompilato, è stato fatto con gli studenti delle scuole superiori. Le stature rilevate sono state contate e sono state trasferite in un grafico a linee per osservare la normale che ne risultava, anche se le osservazioni non erano molte.

Ma il risultato ottenuto non era quello sperato perché la curva non somiglia esattamente ad una distribuzione normale.

Come mai?

Consiglio Nazionale delle Ricerche
Istituto di Ricerche sulla Popolazione e le Politiche Sociali

QUANDO LA SCIENZA È TRASVERSALE
Problem solving e interdisciplinarietà delle competenze matematico - scientifiche a scuola

Spero di aver dato spunti per riflettere ed idee per una didattica efficace

Grazie per l'attenzione

La statistica senza barriere: uso quotidiano (in)consapevole

Loredana Cerbara

loredana.cerbara@irpps.cnr.it

ROMA, 14 maggio 2018