

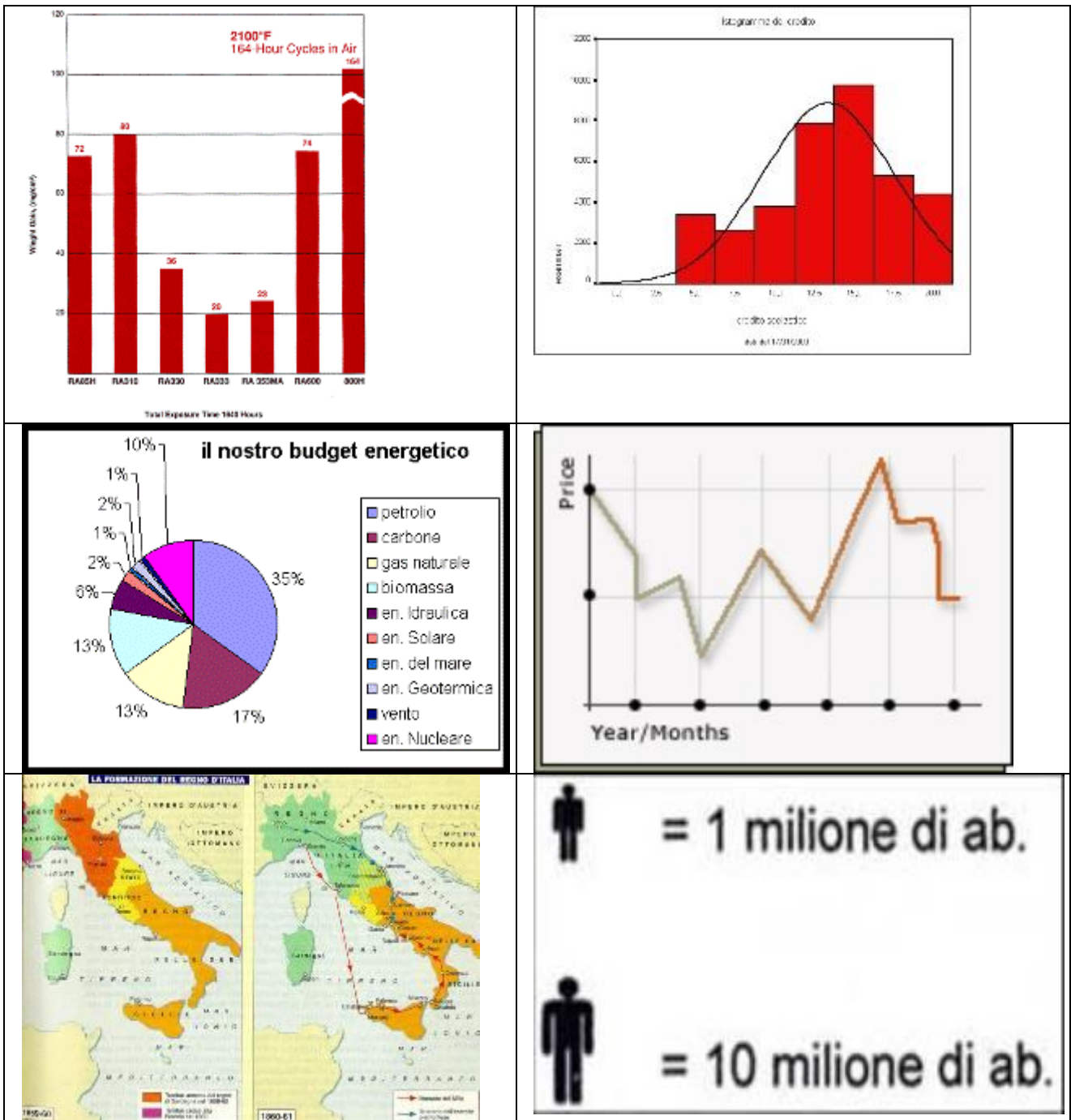
DEFINIZIONI E CONCETTI DI STATISTICA

1. La statistica è quella branca della matematica che si occupa di studiare scientificamente come gli individui di una popolazione sono orientati rispetto ad una certa proposta. Più in generale potremmo dire che essa si occupa dello studio di fenomeni collettivi osservabili nella realtà sociale o in natura o in laboratorio..
2. La POPOLAZIONE è l'insieme degli individui oggetto di un'indagine statistica.
3. Si chiama UNITÀ STATISTICA ciascun individuo della popolazione.
4. In molti casi effettuare un'indagine statistica esaminando tutti gli individui di una popolazione è praticamente impossibile, così si sceglie un sottoinsieme "significativo" della popolazione detto CAMPIONE.
5. La proprietà oggetto dell'indagine statistica si chiama CARATTERE.
6. Un carattere può assumere diversi valori sia numerici sia qualitativi: essi prendono il nome di MODALITÀ.
7. Un carattere che assume modalità espresse da attributi (non numerici) si dice CARATTERE QUALITATIVO.
8. Un carattere descritto da numeri si dice CARATTERE QUANTITATIVO o VARIABILE..
9. Tra i caratteri quantitativi distinguiamo le VARIABILI DISCRETE (espresse da un numero finito di valori numerici o al più da un insieme infinito di valori che può essere messo in corrispondenza biunivoca con l'insieme dei numeri naturali) e le VARIABILI CONTINUE (che possono assumere almeno in linea teorica tutti gli infiniti valori reali di un intervallo).
10. Un'indagine statistica si può suddividere in più fasi:
 - la PIANIFICAZIONE consiste nell'individuare il CARATTERE da studiare e la POPOLAZIONE (o il CAMPIONE) su cui condurre l'indagine;
 - la RILEVAZIONE DEI DATI che nel caso di popolazioni di esseri umani avviene attraverso l'INTERVISTA diretta (un intervistatore pone direttamente delle domande a ogni individuo) o indiretta (facendo compilare un apposito questionario agli individui della popolazione o del campione scelto);
 - l'ELABORAZIONE DEI DATI serve a far emergere da un gran quantitativo di dati alcune informazioni interessanti. Innanzitutto si riordinano i dati e si raggruppano in modo conveniente. È proprio nella fase successiva che si

possono distinguere due tipi di indagini statistiche: la STATISTICA DESCRITTIVA e la STATISTICA INFERENZIALE (o INDUTTIVA). Il primo tipo si occupa di calcolare alcuni numeri significativi per sintetizzare i risultati dell'indagine mentre l'altro tipo ha come obiettivo quello di estendere a tutta la popolazione i risultati ottenuti dall'analisi fatta solo sul campione;

- la PRESENTAZIONE DEI RISULTATI può avvenire attraverso la realizzazione di tabelle, diagrammi o grafici ed ha come obiettivo quello di rendere di immediata lettura i risultati dell'indagine;
 - l'INTERPRETAZIONE DEI RISULTATI è una fase delicata in cui i dati raccolti, classificati e analizzati durante l'indagine vengono elaborati in modo soggettivo per mettere in evidenza alcuni aspetti piuttosto che altri.
11. La FREQUENZA ASSOLUTA di una modalità è il numero di volte che tale modalità si presenta.
 12. La FREQUENZA RELATIVA di una modalità è il rapporto fra la frequenza assoluta della modalità in questione ed il numero totale delle unità statistiche della popolazione (o del campione) considerato.
 13. La FREQUENZA PERCENTUALE altro non è che la frequenza relativa espressa in percentuale.
 14. La FREQUENZA CUMULATA si può definire solo per i caratteri quantitativi: essa rappresenta la somma delle frequenze di tutte le modalità minori o uguali alla modalità cui ci stiamo riferendo.
 15. La SUDDIVISIONE IN CLASSI per una variabile continua consiste nel creare degli intervalli disgiunti.
 16. Il VALORE CENTRALE di una classe è il valore medio fra i due estremi dell'intervallo che identifica la classe.
 17. Le SERIE STATISTICHE sono tabelle in cui la prima colonna è costituita dalle modalità di un carattere qualitativo e la seconda colonna dalle frequenze (assolute o relative) della modalità a cui ci si riferisce.
 18. Le SERIAZIONI STATISTICHE sono tabelle in cui la prima colonna è costituita dalle modalità (discrete o continue) di un carattere quantitativo e la seconda colonna dalle frequenze (assolute o relative) della modalità a cui ci si riferisce.
 19. Le TABELLE A DOPPIA ENTRATA permettono di esaminare contemporaneamente due caratteri diversi relativi alla stessa popolazione.

20. La RAPPRESENTAZIONE GRAFICA dei dati di un'indagine statistica permette una visione d'insieme e un'immediata comprensione della situazione, ma allo stesso tempo se usata in maniera sconveniente o addirittura volutamente fuorviante può trarre in inganno. Ecco le principali rappresentazioni grafiche utilizzate: ORTOGRAMMA, ISTOGRAMMA, AEROGRAMMA, DIAGRAMMA CARTESIANO, CARTOGRAMMA, IDEOGRAMMA:



21. Gli INDICI DI POSIZIONE CENTRALE permettono di cogliere alcuni aspetti importanti del fenomeno. I più utilizzati sono:

- La MEDIA ARITMETICA SEMPLICE

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N}$$

- La MEDIA ARITMETICA PONDERATA

$$\bar{X} = \frac{X_1 \cdot f_1 + X_2 \cdot f_2 + \dots + X_n \cdot f_n}{f_1 + f_2 + \dots + f_n}$$

- La MEDIANA

Se n è dispari allora è il valore centrale;

se n è pari allora è la media aritmetica dei due valori centrali

- La MODA

È il valore a cui corrisponde la frequenza massima

22. Gli INDICI DI VARIABILITÀ servono a descrivere l'attitudine di un fenomeno a manifestarsi sulle varie unità statistiche con modalità diverse e distanti fra loro. I più utilizzati sono:

- Il CAMPO DI VARIAZIONE

$$X_N - X_1$$

- Gli SCARTI ASSOLUTI

$$|X_1 - \bar{X}|, |X_2 - \bar{X}|, \dots, |X_N - \bar{X}|$$

- Lo SCARTO SEMPLICE MEDIO

$$s = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n}$$

- Gli SCARTI QUADRATICI

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$$

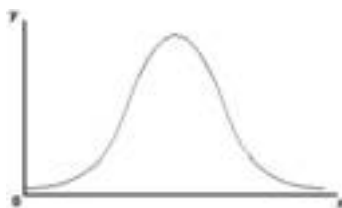
- La VARIANZA

$$V = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

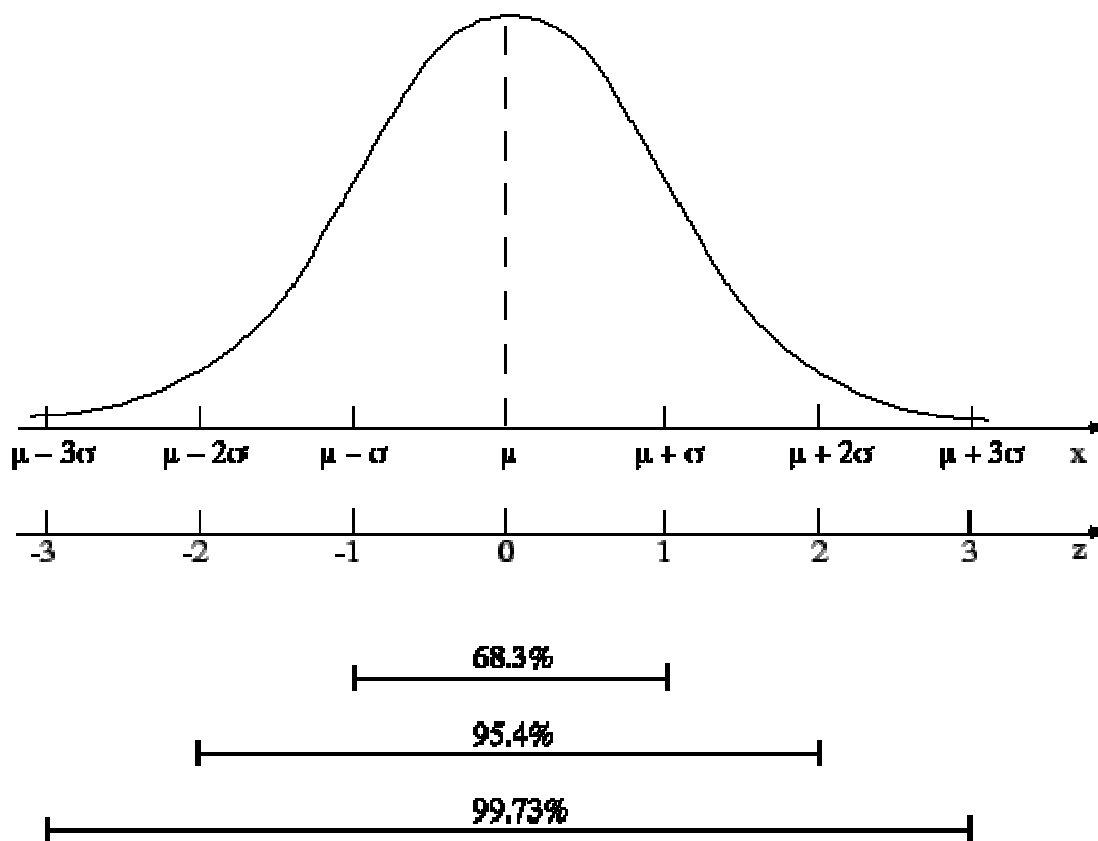
- Lo SCARTO QUADRATICO MEDIO o DEVIAZIONE STANDARD

$$\sigma = \sqrt{V} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

23. La **DISTRIBUZIONE GAUSSIANA** è una distribuzione di frequenze che da luogo ad una curva particolare detta **CURVA NORMALE** o **GAUSSIANA** dalla caratteristica forma “a campana”.



24. **PROPRIETÀ** della **CURVA DI GAUSS**. Il calcolo della deviazione standard ha particolare importanza nelle distribuzioni gaussiane perché in tali distribuzioni, se chiamiamo M la media aritmetica di distribuzione e σ la sua deviazione standard, allora il 68,3% dei valori è compreso fra $M - \sigma$ e $M + \sigma$, il 95,4% fra $M - 2\sigma$ e $M + 2\sigma$ e addirittura il 99,73% fra $M - 3\sigma$ e $M + 3\sigma$.



25. Nella statistica induttiva o inferenziale, i dati statistici relativi al campione forniscono una stima dei corrispondenti dati statistici relativi a tutta la popolazione. Certo è che identificando la media della popolazione con la media del campione si commette inevitabilmente un errore di approssimazione. L'**ERRORE STANDARD** valuta questa incertezza e si definisce come

$$s_x = \frac{s}{\sqrt{n-1}} \quad \text{dove } s \text{ è la deviazione standard relativa al campione.}$$

26. Per stimare il valore della media della popolazione si può individuare un intervallo che lo contenga con una certa probabilità. Si chiama **INTERVALLO DI CONFIDENZA** o **FORBICE** l'intervallo che ha per estremi il valore medio del campione diminuito del triplo dell'errore standard ed il valore medio aumentato del triplo dell'errore standard, cioè: $[\bar{x} - 3\sigma, \bar{x} + 3\sigma[$ ed esso rappresenta un intervallo in cui cadrà la media della popolazione con una percentuale del 99,74%.

Si procede allo stesso modo se si vuole individuare l'**INTERVALLO DI CONFIDENZA** di una certa caratteristica della popolazione. Ad esempio se f è la frequenza percentuale di una certa modalità rilevata su un campione di n unità statistiche, allora l'errore standard si calcola con la seguente formula:

$$s_f = \sqrt{\frac{f \cdot (1 - f)}{n}}$$

e ancora una volta l'intervallo di confidenza sarà quello che ha per estremi il valore della percentuale f diminuito del triplo dell'errore standard ed il valore della percentuale f aumentato del triplo dell'errore standard, cioè

$$|f - 3s_f, f + 3s_f|$$