Prerequisiti:

- Primi elementi di probabilità e statistica.
- Nozioni di calcolo combinatorio.
- Rappresentazione di punti e rette in un piano cartesiano.
- Concetto di distribuzione di probabilità
- Conoscenza della distribuzione normale e delle sue caratteristiche

Questa unità interessa tutte le scuole superiori, ad eccezione del Liceo Artistico, ma con modalità diverse, che saranno specificate a tempo debito.

OBIETTIVI DI APPRENDIMENTO

Una volta completata l'unità, gli allievi devono essere in grado di:

- *spiegare lo scopo della statistica inferenziale e, in particolare, in cosa consiste un ragionamento induttivo*
- *spiegare come si costruisce un campione rappresentativo di una popolazione*
- *spiegare le caratteristiche di una distribuzione campionaria*
- *spiegare in cosa consiste il test delle ipotesi e come si articola*
- *formalizzare e risolvere semplici problemi di verifica delle ipotesi, trattabili con il ricorso alla distribuzione normale, relativi alla media ed alla proporzione di un collettivo statistico, utilizzando se del caso strumenti di calcolo automatico*
- *spiegare i concetti di "livello di confidenza", "intervallo di confidenza", "errore massimo"*
- *stimare entro quali intervalli è compresa la misura di una grandezza della quale sono state effettuate misure sperimentali in numero adeguato, utilizzando eventualmente uno strumento di calcolo automatico*
- *conoscere e applicare il teorema centrale del limite.*

Cenni di statistica inferenziale

Unità 81

81.1 IL RAGIONAMENTO INDUTTIVO [1]

81.1.1 Hai imparato dallo studio della statistica descrittiva che la tabulazione di una serie di dati statistici e la loro rappresentazione grafica hanno il pregio di descrivere in modo ordinato e completo il fenomeno sul quale s'indaga.

Hai pure imparato che a volte è comodo, e spesso anche necessario, sintetizzare con un solo indice l'andamento del fenomeno, a condizione naturalmente che questo valore di sintesi sia idoneo a riassumere le caratteristiche del collettivo che interessa evidenziare. L'indice più frequente è la *media aritmetica*.

Assieme alla media aritmetica – che è un *indice di posizione* – hai imparato a considerare anche alcuni *indici di dispersione*, come la *deviazione standard*, la quale indica come i dati del collettivo considerato si disperdono intorno alla media aritmetica.

Attraverso lo studio della probabilità hai incominciato ad imparare, inoltre, un fatto notevole: è possibile ottenere informazioni sulle caratteristiche di una popolazione statistica ragionando solo su un sottoinsieme di individui estratti da essa in modo da rappresentarla e detto più propriamente **campione**.

Beninteso, non si tratta di informazioni certe, del tipo "se questa è la misura del lato di un quadrato, allora questa è la sua area". Si tratta, invece, di informazioni di tipo probabilistico, del tipo "questa è la misura della caratteristica del campione, per cui la misura della caratteristica corrispondente del collettivo sarà quest'altra con questa probabilità e questo margine di errore".

Di solito, le misure delle caratteristiche dell'intero collettivo sono indicate con il nome di ***parametri***, mentre quelle riguardanti un campione sono indicate con il nome di ***statistiche***. Anche i simboli, usati per indicare le misure di tali caratteristiche, si assumono di solito diversi, a seconda che siano riferite al collettivo o al campione. Lo vedremo strada facendo.

I *parametri* sono delle costanti per una determinata popolazione, ma in genere non sono conosciuti, specialmente se la popolazione è molto numerosa (diciamo dell'ordine della decina di migliaia).

Le *statistiche* variano da campione a campione, ma, una volta che è stato individuato il campione, si possono calcolare e perciò sono valori noti.

Il fatto importante è che, una volta che siano conosciute le statistiche di un campione, si possono ***stimare*** i parametri corrispondenti della popolazione. Solo "stime" di tali parametri però, e perciò inficiate da errori, e non conoscenza certa dei loro effettivi valori.

Il modo di ragionare che, partendo dalle statistiche di un campione rappresentativo di una popolazione, permette di stimare i parametri dell'intera popolazione, si chiama **ragionamento induttivo**.

81.1.2 Il "ragionamento induttivo" è alla base della **statistica inferenziale** (o **inferenza statistica** o **induzione statistica** o **statistica induttiva** o **statistica matematica**).

La statistica inferenziale insegna a stabilire:

a) come ottenere un campione di una popolazione statistica che sia effettivamente rappresentativo dell'intera popolazione;

b) di quanti individui deve essere composto il campione, in rapporto a quelli che compongono la popolazione su cui si indaga e in base all'errore massimo che si ipotizza di commettere nella valutazione dei risultati e al grado di affidabilità di questi risultati;

c) quali relazioni sussistono fra le caratteristiche del campione (le *statistiche*) e quelle corri-

[1] Questo paragrafo interessa tutte le scuole: Tecnici e Professionali nel 2° biennio, Licei nella 5ª classe.

spondenti del collettivo (i *parametri*);

d) con quale grado di probabilità e con quale margine di errore sono inclusi in un certo intervallo la media di un collettivo, se si indaga su una caratteristica quantitativa (per esempio le altezze di un gruppo di persone), oppure la percentuale di un universo statistico, se si indaga su una caratteristica qualitativa (per esempio coloro che rispondono affermativamente ad una certa domanda); ammesso ovviamente di conoscere media (o percentuale) e deviazione standard del campione rappresentativo del collettivo.

Di tutto ciò non possiamo occuparci in maniera approfondita per mancanza di adeguati strumenti matematici. Possiamo, tuttavia, utilizzare i risultati ai quali la teoria permette di pervenire, anche senza dimostrarli, quantomeno per avere un'idea di ciò che concretamente fa l'inferenza statistica.

Il **metodo induttivo** – come forma di ragionamento che passa dal particolare all'universale – ha in **Aristotele** (384-322 a.C.) il primo studioso, anche se questo grande filosofo ne attribuisce la paternità a **Socrate** (470-399 a.C.). Aristotele, comunque, non annetteva al metodo induttivo capacità di affermazioni logicamente vincolanti e questa concezione durò a lungo, addirittura fino al XVII secolo. Il primo a metterla in discussione fu il filosofo inglese **Francis Bacon** (1561-1636) con la pubblicazione, avvenuta nel 1620, dell'opera ***Novum Organum***. Le sue idee furono riprese, ma oltre due secoli più tardi, dal filosofo ed economista inglese **John Stuart Mill** (1806-1873), il quale, nell'opera ***A System of Logic*** (1843), giunse alla conclusione che i ragionamenti induttivi, pur legittimi, non portano però a leggi "certe" ma solo "probabili".

È il concetto che sta alla base dell'inferenza statistica, ma per trovarne concrete applicazioni bisogna aspettare il contributo di altri studiosi. Uno di questi fu il matematico e statistico inglese **Karl Pearson** (1857-1936). Ma è un suo discepolo, l'inglese **Ronald Aylmer Fisher** (1890-1962), docente di genetica, che è considerato il vero fondatore della moderna statistica inferenziale. Fisher fu tra i primi, se non addirittura il primo, a comprendere che alla base dell'inferenza c'era il campionamento casuale.

È doverosa una precisazione. Se è vero che inizialmente Fisher si formò anche, ma non solo, sui lavori di Pearson, è ugualmente vero che ben presto si staccò dalle idee di questi, dal quale si rese assolutamente indipendente. Questo fu causa di una forte ostilità tra i due, che praticamente non si placò mai.

Due altri personaggi meritano di essere citati per i notevoli e fondamentali contributi allo sviluppo della statistica: si tratta del polacco **Jerzy Neyman** (1894-1981) e del britannico **Egon Sharpe Pearson** (1895-1980), figlio di Karl.

81.2 CAMPIONAMENTO E MEDIA CAMPIONARIA [2]

81.2.1 Affinché un campione di una popolazione sia effettivamente rappresentativo della popolazione su cui s'indaga è necessario che sia assolutamente casuale, in modo che tutti gli individui del collettivo abbiano la stessa probabilità di essere inclusi nel campione.

- ESEMPIO. Fra i 653 dipendenti di una ditta si vuole scegliere un campione casuale di 10 soggetti. Per questo si scrivono i nomi dei 653 dipendenti su altrettanti foglietti di carta e si inseriscono in un'urna.

[2] Anche questo paragrafo riguarda tutte le scuole: Tecnici e Professionali nel 2° biennio, Licei nella 5ª classe.

Quindi si procede all'estrazione casuale di 10 foglietti: i corrispondenti dipendenti costituiscono il campione casuale dei 10 soggetti.

È questo un metodo certamente rudimentale e grossolano, che però fa capire bene quello che vogliamo dire quando affermiamo che il campione deve essere assolutamente casuale.

Di fatto, esistono tecniche più complesse e raffinate, in cui svolgono un ruolo fondamentale i cosiddetti *numeri casuali*, ma riteniamo che non sia il caso di occuparcene. Avrà modo di approfondire la questione chi proseguirà gli studi in settori disciplinari in cui la Statistica svolge un ruolo importante.

81.2.2 Il numero di individui che compongono un campione rappresentativo di una popolazione (o la popolazione stessa) si chiama ***numerosità*** o ***dimensione*** del campione (o della popolazione).

Supponiamo di estrarre da una popolazione di dimensione N tutti i possibili campioni di dimensione n, con $n<N$ ovviamente. Ammettiamo che ogni unità del campione, una volta estratta, non sia rimessa nella popolazione [3]: si parla di ***campionamento senza reinserimento***. Il numero k di tali campioni è uguale al numero delle combinazioni semplici delle N unità che compongono la popolazione prese ad n ad n; vale a dire:

$$k = \binom{N}{n}.$$

- Per esempio, se la popolazione è costituita dalle seguenti unità:

$$x_1,\ x_2,\ x_3,\ x_4,$$

aventi tutte la stessa probabilità di essere estratte, tutti i campioni di dimensione 2, estraibili da essa (senza reinserimento), sono i seguenti insiemi:

$$\{x_1, x_2\},\ \{x_1, x_3\},\ \{x_1, x_4\},\ \{x_2, x_3\},\ \{x_2, x_4\},\ \{x_3, x_4\},$$

e sono evidentemente in numero di $\binom{4}{2}$.

Si capisce che, se si estrae a sorte fra tali k campioni, ognuno di essi avrà probabilità $1/k$ di essere estratto. Per cui le probabilità $p_1, p_2, \dots, p_k$ di estrazione dei k campioni sono tali che:

$$\sum_{i=1}^{k} p_i = \underbrace{\frac{1}{k}+\frac{1}{k}+\dots+\frac{1}{k}}_{\text{k addendi}} = 1\,.$$

Consideriamo ora le medie dei k campioni suddetti: $\mu_1, \mu_2, \dots, \mu_k$.

Sono dette ***medie campionarie*** e in genere sono diverse fra loro.

- Ritornando all'esempio considerato prima, le $k=6$ medie dei campioni estratti sono:

$$\mu_1=\frac{x_1+x_2}{2},\quad \mu_2=\frac{x_1+x_3}{2},\quad \mu_3=\frac{x_1+x_4}{2},\quad \mu_4=\frac{x_2+x_3}{2},\quad \mu_5=\frac{x_2+x_4}{2},\quad \mu_6=\frac{x_3+x_4}{2}.$$

Poiché $\sum_{i=1}^{k} p_i = 1$, la variabile $\overline{X}$, rappresentata dalla seguente matrice:

$$\overline{X} = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_k \\ p_1 & p_2 & \dots & p_k \end{bmatrix}$$

è una variabile aleatoria. Si può prendere perciò in considerazione la sua distribuzione di probabilità: si chiama **distribuzione della media campionaria**.

Questa distribuzione ha una sua media aritmetica e una sua deviazione standard, che si indicano con i simboli rispettivamente: $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$.

[3] Un campione ottenuto **con reinserimento** delle unità estratte si definisce ***bernoulliano***. La denominazione deriva dal matematico svizzero Jakob Bernoulli (1654-170).

- Ritornando ancora una volta all'esempio già considerato, riguardo alla media $\mu_{\overline{X}}$ si ha:

$$\mu_{\overline{X}}=\frac{\mu_1+\mu_2+\mu_3+\mu_4+\mu_5+\mu_6}{6}=\frac{\frac{x_1+x_2}{2}+\frac{x_1+x_3}{2}+\frac{x_1+x_4}{2}+\frac{x_2+x_3}{2}+\frac{x_2+x_4}{2}+\frac{x_3+x_4}{2}}{6}=$$

$$=\frac{3(x_1+x_2+x_3+x_4)}{12}=\frac{x_1+x_2+x_3+x_4}{4}=\mu,$$

dove μ è evidentemente la media aritmetica della popolazione da cui sono estratti i campioni.
Sarebbe possibile dimostrare che questa stessa relazione, $\mu_{\overline{X}}=\mu$, vale per ogni dimensione N della popolazione ed n dei campioni. La dimostrazione non differisce sostanzialmente da quella da noi proposta nel caso particolare in cui $N=4$ ed $n=2$, ma non ce ne occuperemo.
Per quanto riguarda il calcolo di $\sigma_{\overline{X}}$, le cose sono un po' più complicate. Ci limitiamo a fornire il risultato, che quando la dimensione n dei campioni è "sufficientemente piccola" rispetto alla dimensione N della popolazione (in pratica quando $n<0{,}05N$) e, lo ripetiamo, quando l'estrazione è effettuata senza reinserimento, è il seguente:

$$\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}$$

dove σ è la deviazione standard della popolazione da cui sono estratti i campioni.
Anche se noi, negli esempi presi in considerazione, abbiamo presentato per comodità situazioni in cui le dimensioni N ed n sono poco diverse l'una dall'altra, la circostanza che n sia "piccola" rispetto ad N è assai frequente nelle indagini statistiche: si pensi, ad esempio, al corpo elettorale italiano (costituito da decine di milioni di soggetti) e ad un campione di elettori (formato in genere da qualche migliaio di unità); oppure, altro esempio, ai potenziali acquirenti di un certo prodotto (svariate migliaia di persone) e ad un campione su cui si vuole indagare (qualche centinaio di soggetti).

Il fatto che la distribuzione della media campionaria di campioni casuali della medesima dimensione n estratti da una popolazione qualsiasi (di varianza finita) si approssimi alla distribuzione normale con media μ e deviazione standard $\sigma/\sqrt{n}$, quando n è sufficientemente grande ($n\geq 30$), è una proprietà che va sotto il nome di ***Teorema centrale del limite*** [(4)]. Naturalmente μ e σ sono la media e la deviazione standard della popolazione.

81.2.3 In realtà, la formula corretta per il calcolo di $\sigma_{\overline{X}}$ – sempre che l'estrazione avvenga senza reinserimento – è la seguente:

$$\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}.$$

Non avremo occasione di servircene. Ciò non di meno, ti proponiamo un esercizio idoneo a verificare tale formula, almeno in un caso particolare.

Si consideri la popolazione costituita dalle seguenti unità statistiche: 2, 4, 6, 8, aventi tutte la stessa probabilità di essere estratte. Di tale popolazione calcolare la deviazione standard σ.
Una volta estratti da essa tutti i possibili campioni di dimensione 2, calcolare la deviazione standard $\sigma_{\overline{X}}$ della distribuzione della media campionaria sia direttamente sia utilizzando la formula suddetta. Verificare che

[4] Il teorema centrale del limite fu dimostrato, nel 1922, dal matematico e statistico finlandese **Jan Waldemar Lindeberg** (1876-1932) e successivamente, in modo indipendente, dal britannico **Alan Turing** (1912-1954), uno dei padri dell'informatica.

i due risultati coincidono.

La formula in questione, quando $N \to \infty$, considerato che il radicale che in essa figura tende ad 1, si identifica con la precedente:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

NOTA BENE: Quest'ultima formula vale pure quando l'estrazione avviene con reinserimento, anche da una popolazione finita, potendosi in tal caso l'estrazione assimilare all'estrazione senza reinserimento da una popolazione di dimensione N infinita.

81.2.4 Supponiamo, allora, che si debba indagare su una popolazione di dimensione N, media μ e deviazione standard σ. Si suppone di estrarre da essa, senza reinserimento, tutti i possibili campioni casuali della stessa dimensione n, e siano $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ la media e la deviazione standard della distribuzione della media campionaria. Com'è già stato precisato, risulta in ogni caso:

[1] $$\mu_{\overline{X}} = \mu\ ;$$

mentre solo se n è sufficientemente piccolo rispetto ad N (in pratica se $n<0{,}05N$) si ha, con buona approssimazione:

[2] $$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Tutto questo se l'indagine riguarda un aspetto quantitativo, come per esempio una lunghezza, un peso, eccetera.

Supponiamo, invece, che l'indagine riguardi la presenza o meno di un dato carattere nella popolazione, come, tanto per fare degli esempi, il gradimento di un certo prodotto o la dichiarazione di voto per un determinato partito. Indichiamo, allora, con p la percentuale della popolazione che ha quel carattere e con σ la relativa deviazione standard e indichiamo con $\mu_{\overline{p}}$ e $\sigma_{\overline{p}}$ la percentuale e la deviazione standard della distribuzione della media campionaria. Risulta, in ogni caso:

[1'] $$\mu_{\overline{p}} = p\ ,$$

mentre solo se $n<0{,}05N$ si ha:

[2'] $$\sigma_{\overline{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Ora, nella pratica non si prendono in esame tutti i possibili campioni estraibili da una popolazione e nemmeno alcuni di essi, ma si opera su un solo campione. Se questo ha dimensione n, media $\overline{x}$ (o, eventualmente, percentuale f) e deviazione standard s, si assumono come ***stime*** di $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ (eventualmente $\mu_{\overline{p}}$ e $\sigma_{\overline{p}}$) gli stessi valori forniti dalle [1] e [2] (o rispettivamente dalle [1'] e [2']) ponendo, di fatto, $\overline{x}$ al posto di μ, f al posto di p ed s al posto di σ. Per cui si ha:

[3] $$\mu_{\overline{X}} = \overline{x} \quad \text{e} \quad \sigma_{\overline{X}} = \frac{s}{\sqrt{n}}$$

oppure eventualmente:

[3'] $$\mu_{\overline{p}} = f \quad \text{e} \quad \sigma_{\overline{p}} = \sqrt{\frac{f(1-f)}{n}}.$$

- Riassumendo:

a) per avere informazioni sulla media μ di una popolazione, bisogna conoscere la media $\bar{x}$ di un campione casuale rappresentativo della popolazione.

b) per avere informazioni sulla percentuale p di una popolazione, relativa ad un dato carattere, bisogna conoscere la percentuale f di un campione casuale rappresentativo della popolazione.

S'intende che la media (o la percentuale) del campione non è esattamente uguale a quella della popolazione, ma è solo una **stima** di essa. Si commette, perciò, un errore assumendo la media del campione in sostituzione della media della popolazione, come pure la percentuale del campione che presenta un dato carattere in sostituzione della percentuale della popolazione che presenta quel carattere. Precisamente, indicando con ε quest'errore, si ha:

$$\mu = \bar{x} \pm \varepsilon \quad \text{oppure:} \quad p = f \pm \varepsilon.$$

Sorgono, allora, un paio di interrogativi:

a) quanto vale l'errore che si commette? b) l'errore è lo stesso per qualunque campione?

La teoria permette di rispondere ad entrambe le domande. Vedremo come fra breve.

81.2.4 In sintesi, lo studio e l'analisi dei fenomeni statistici sono condotti mediante la formulazione di un modello teorico e la verifica della sua rispondenza ai dati reali. Alcuni degli strumenti statistici più usati per tale analisi sono la ***correlazione***, la ***regressione lineare***, il ***test delle ipotesi***, la ***stima*** dei parametri di un collettivo statistico sulla base delle osservazioni campionarie.

Per quanto attiene alla correlazione ed alla regressione lineare rimandiamo alle nozioni esposte a suo tempo sull'argomento (cfr.: Unità 55).

Riguardo al test delle ipotesi ed al problema della stima ce ne occuperemo brevemente in questa unità di qui a poco, segnalando che la comprensione di entrambi gli argomenti presuppone come requisito indispensabile l'avere appreso le caratteristiche fondamentali della distribuzione normale, di cui ci siamo occupati nella precedente unità.

Prima di passare al test delle ipotesi ed al problema della stima vogliamo però accennare ad un altro impiego della distribuzione normale, il cosiddetto "controllo di qualità".

81.3 CONTROLLO DI QUALITÀ [5]

80.3.1 La distribuzione di Gauss trova applicazione nel controllo di qualità dei servizi e dei prodotti. Una trattazione esauriente dell'argomento non è possibile, per cui dobbiamo accontentarci di un breve accenno.

81.3.2 Un'azienda vuole stimare se il processo di lavorazione di determinati oggetti (per esempio: bulloni di dato diametro) produce pezzi da scartare perché difettosi (nell'esempio dei bulloni: diametro troppo piccolo o troppo grande). Per questo prende un campione casuale dei pezzi prodotti e calcola la media aritmetica μ della grandezza interessata (nell'esempio: media dei diametri) e la deviazione standard σ. L'azienda fissa quindi la *tolleranza*, vale a dire la distanza τ dalla media μ. Ammesso ora che la distribuzione del campione sia assimilabile alla distribuzione normale (dimensione del campione non inferiore a 30) e indicata con G_C la grandezza dei pezzi conformi (cioè non difettosi) del campione, si as-

[5] Questo paragrafo riguarda soltanto Tecnici e Professionali, che ne affronteranno lo studio nel 2° biennio.

sume quanto segue:

- Il processo genera pezzi di scarto se $\tau<3\sigma$, per cui si ha:
$$\mu-3\sigma<\mu-\tau<G_C<\mu+\tau<\mu+3\sigma.$$
- Il processo non genera pezzi di scarto, ma non presenta margini di manovra, se $\tau=3\sigma$, per cui si ha:
$$\mu-3\sigma=\mu-\tau<G_C<\mu+\tau=\mu+3\sigma.$$
- Il processo non genera pezzi di scarto e vi sono inoltre margini di manovra se $\tau>3\sigma$, per cui si ha:
$$\mu-\tau<\mu-3\sigma<G_C<\mu+3\sigma<\mu+\tau.$$

Un esempio per chiarire.

- ESERCIZIO. Un'azienda produce funi d'acciaio aventi un carico di rottura di $1570\ N/mm^2$ (Newton su millimetri quadrati) con una tolleranza del 4%. L'azienda, al fine di stimare se il processo di lavorazione genera pezzi di scarto, estrae un campione casuale di 30 funi e di ciascuna di esse misura il carico di rottura. I risultati ottenuti sono riportati in una tabella (Tab. 1). A quale conclusione giunge l'azienda?

TAB. 1 – Carico di rottura delle funi del campione (in N/mm^2)

1560	1490	1632	1504	1597	1644	1635	1596	1598	1650
1613	1588	1609	1574	1496	1537	1529	1630	1601	1563
1559	1637	1624	1545	1556	1577	1606	1528	1517	1520

RISOLUZIONE. Si calcola anzitutto la tolleranza τ, la media aritmetica μ del campione e la deviazione standard σ. Si trova (*l'uso di un foglio elettronico rende facile il compito*):
$$\tau = 62{,}8\,;\quad \mu \approx 1577\,;\quad \sigma \approx 46\,.$$
Risulta evidentemente: $\tau<3\sigma$, per cui il processo di lavorazione produce pezzi di scarto e va migliorato.

81.3.3 Il controllo, invece della misura di una grandezza, può interessare la presenza di un determinato carattere. Le cose vanno alla stessa maniera con le solite differenze: la percentuale f del campione che presenta quel carattere al posto della media μ delle grandezze; la percentuale P_C dei pezzi conformi (cioè con quel carattere) al posto della grandezza G_C.

Ancora un esempio a chiarimento di tutto ciò.

- ESERCIZIO. Un'azienda produce viti d'acciaio con filettatura a profilo triangolare con angolo di profilo di 60°. La tolleranza di viti difettose è del 4%. L'azienda, al fine di stimare se il processo di lavorazione delle viti produce pezzi di scarto, estrae un campione casuale di 300 viti e trova che 290 di esse sono conformi. A quale conclusione giunge?

RISOLUZIONE. Si constata anzitutto che:
$$\tau=0{,}04\,;\quad f=\frac{290}{300}\,;\quad \sigma=\sqrt{\frac{f(1-f)}{n}}=\sqrt{\frac{\frac{290}{300}\cdot\left(1-\frac{290}{300}\right)}{300}}\approx 0{,}010\,.$$
Siccome $3\sigma\approx 0{,}03$ risulta evidentemente che $\tau>3\sigma$, per cui non solo il processo di lavorazione non produce pezzi di scarto ma addirittura ci sono margini di manovra per modificarlo convenientemente in modo da risparmiare sulla produzione.

81.4 IL TEST DELLE IPOTESI [6]

81.4.1 Nello studio e nell'analisi di un fenomeno statistico, una volta formulato un modello teorico, si tratta di valutare se lo stesso è compatibile con i dati reali. Ciò è fatto mediante il **test delle ipotesi**, detto anche **verifica delle ipotesi**.

Ecco come si procede di norma:

- si formula un'ipotesi circa un determinato fenomeno (la formulazione di un modello teorico è già un'ipotesi);
- si studia un carattere del fenomeno mediante un campione casuale rappresentativo del fenomeno medesimo;
- sulla base del carattere rilevato nel campione l'ipotesi formulata è accettata o rifiutata.

L'ipotesi da verificare è chiamata **ipotesi nulla** ed è indicata con $\mathbf{H_0}$, mentre l'**ipotesi alternativa** è indicata con $\mathbf{H_1}$.

Si capisce che, trattandosi di valutazioni statistiche, può capitare di scartare un'ipotesi che invece è buona o accettare un'ipotesi che al contrario è pessima. Tecnicamente si dice che si possono commettere due tipi di errore:

- ***errore di prima specie*** quando si rifiuta un'ipotesi che invece è vera,
- ***errore di seconda specie*** quando si accetta un'ipotesi falsa.

La probabilità di rifiutare un'ipotesi quando è vera, vale a dire la probabilità di commettere un errore di prima specie è indicata con $\boldsymbol{\alpha}$ e si chiama anche **livello di significatività**. Naturalmente, $\mathbf{1-\alpha}$ è la probabilità di accettare un'ipotesi quando è vera: si chiama **livello di confidenza**.

Invece, la probabilità di accettare un'ipotesi quando è falsa, vale a dire la probabilità di commettere un errore di seconda specie è indicata con $\boldsymbol{\beta}$. Il numero $\mathbf{1-\beta}$ è evidentemente la probabilità di rifiutare un'ipotesi falsa: si chiama **potenza del test**.

La probabilità di commettere un errore di prima specie è nota nello stesso momento in cui è fissato il livello di significatività di un test. Diversamente il calcolo della probabilità di commettere un errore di seconda specie richiede la conoscenza di altri dati. Considerato il livello elementare di questa trattazione, qui ci occupiamo solamente del livello di significatività (e conseguentemente del livello di confidenza), lasciando a studi universitari la valutazione della potenza di un test.

Forniamo, ad ogni buon conto, una tabella (Tab. 2) che riassume le situazioni che possono presentarsi a seconda che si rifiuti o si accetti l'ipotesi nulla, che può essere vera o falsa.

Ipotesi nulla H_0 **Decisione**	**VERA**	**FALSA**
Rifiuto di H_0	Errore di prima specie. Probabilità = livello di significatività = α	Decisione corretta. Probabilità = potenza del test = $1-\beta$
Accettazione di H_0	Decisione corretta. Probabilità = livello di confidenza = $1-\alpha$	Errore di seconda specie. Probabilità = β

TAB. 2

Al fine di tenere bassa la probabilità di commettere un errore di prima specie, il livello di significatività più comunemente usato dagli statistici è quello del 5%, ma non è l'unico. Sono infatti usati altri due

[6] Anche questo paragrafo riguarda il 2° biennio di Tecnici e Professionali. Ma riguarda pure la 5ª classe del Liceo delle Scienze umane, opzione Economico-sociale.

livelli: quello dell'1% e quello del 10%. Il primo abbastanza spesso, il secondo molto raramente. Quando è usato il livello di significatività dell'1% si dice anche che il test è *molto significativo*. Quando è usato il livello del 10% si dice a volte che il test è *poco significativo*.
Naturalmente, in corrispondenza dei seguenti livelli di significatività:

$$1\%,\ \ 5\%,\ \ 10\%,$$

si hanno i seguenti livelli di confidenza:

$$99\%,\ \ 95\%,\ \ 90\%.$$

ATTENZIONE! Il livello di significatività NON è la probabilità che l'ipotesi nulla sia falsa, così come il livello di confidenza NON è la probabilità che essa sia vera. Il primo è semplicemente la probabilità di rifiutare l'ipotesi nulla (che, se è vera, implica un errore di prima specie), ed il secondo la probabilità di accettarla (che, sempre che sia vera, è una decisione corretta).

81.4.2 Ci occuperemo del test delle ipotesi relativamente alle due seguenti caratteristiche:
- media di un collettivo statistico;
- proporzione di un collettivo statistico.

Lo faremo testando campioni di dimensione $n \geq 30$, in modo da poter fare ricorso alla distribuzione normale. Più precisamente alla distribuzione normale standardizzata, la cui equazione è, come noto, la seguente:

$$\mathbf{y = \frac{1}{2\pi}\, e^{-x^2/2}}.$$

E lo faremo prendendo in considerazione due esempi particolari.

- ESERCIZIO 1. *Test delle ipotesi relativo alla media di un collettivo statistico*.

Un'impresa produce lampadine elettriche. L'ipotesi è che esse durino mediamente 1.200 ore. L'impresa intende effettuare un test di verifica di tale ipotesi al livello di significatività del 5% e, a tal fine, prende un campione casuale di 90 lampadine e, fatte le debite indagini e le necessarie misurazioni, trova una media di durata di 1180 ore con una deviazione standard di 110 ore. Quali conclusione si traggono?
RISOLUZIONE. Premettiamo alcune considerazioni che, quantunque riferite alla situazione specifica, hanno tuttavia carattere generale.
Indichiamo con μ la media di durata delle lampadine prodotte dall'impresa e con μ_0 la media di durata delle lampadine del campione preso in esame.
Si adottano le due seguenti ipotesi:
- <u>ipotesi nulla</u> (H_0): $\mu = \mu_0$;
- <u>ipotesi alternativa</u> (H_1): $\mu \neq \mu_0$.

Al livello di significatività del 5%, e di conseguenza al livello di confidenza del 95%, la statistica ci insegna che la regione di accettazione dell'ipotesi nulla è compresa fra i valori $-1{,}96$ e $+1{,}96$. Per questo motivo la regione sottostante alla curva normale standardizzata compresa fra $-1{,}96$ e $+1{,}96$ si chiama **regione di accettazione** dell'ipotesi nulla. Invece la regione sottostante alla stessa curva, ma esterna alla regione di accettazione è chiamata **regione del rifiuto** dell'ipotesi nulla (Fig. 1).

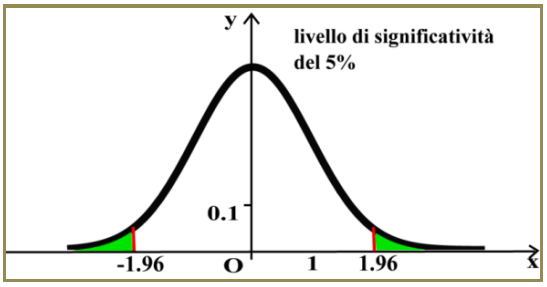

FIG. 1 – Test bilaterale

In considerazione del fatto che la regione del rifiuto è costituita dalle due regioni periferiche sotto la curva normale standardizzata, il test delle ipotesi è chiamato ***test a due code*** (o ***bilaterale***).

A questo punto bisogna convertire nella variabile z standardizzata la media campionaria e controllare se risulta $z<-1{,}96$ o $z>+1{,}96$: se così è, considerato che z cade nella regione del rifiuto dell'ipotesi nulla, quest'ipotesi va appunto rifiutata. Se invece è $-1{,}96<z<+1{,}96$ l'ipotesi nulla va accettata.

Ricordiamo che la formula idonea a standardizzare la media campionaria è la seguente (cfr.: Unità 80, paragrafo n. 80.1.3, dove al posto di X, x, σ scriviamo adesso rispettivamente z, $\mu_{\overline{X}}$, $\sigma_{\overline{X}}$):

[4] $$z = \frac{\mu_{\overline{X}} - \mu}{\sigma_{\overline{X}}}$$

essendo μ la media del collettivo, $\mu_{\overline{X}}$ la media campionaria e $\sigma_{\overline{X}}$ la deviazione standard campionaria.

Nel nostro caso, in cui μ_0 è la media del campione preso in esame ed s la deviazione standard di tale campione, si ha:

$$\mu_{\overline{X}}=\mu_0=1180 \text{ ore}, \quad \sigma_{\overline{X}}=\frac{s}{\sqrt{n}}=\frac{110}{\sqrt{90}} \text{ ore}, \quad z=\frac{1180-1200}{\frac{110}{\sqrt{90}}} \approx -1{,}72\,.$$

Poiché il valore di z così trovato non appartiene alla regione del rifiuto dell'ipotesi nulla, ma a quella dell'accettazione ($-1{,}72>-1{,}96$), l'impresa, al livello di significatività del 5%, deve accettare tale ipotesi e concludere che effettivamente la durata delle lampadine si può ritenere mediamente uguale a 1200 ore.

Si può constatare che se il livello di significatività fosse stato del 10%, la regione del rifiuto sarebbe stata esterna alla striscia compresa fra $-1{,}65$ e $+1{,}65$. In questo caso essendo $z=-1{,}72<-1{,}65$ l'ipotesi nulla va rifiutata e va accettata invece l'ipotesi alternativa. In altri termini l'impresa, al livello di significatività del 10%, deve rifiutare l'ipotesi che la durata delle lampadine sia mediamente di 1200 ore.

- ESERCIZIO 2. *Test delle ipotesi relativo alla proporzione di un collettivo statistico*.

Sugli iscritti alla prima classe dell'Istituto "Einstein" prima dell'anno 2006, il 92,5% ha conseguito il diploma entro la durata regolare di 5 anni di liceo. Nell'anno 2006 si sono iscritti alla prima classe dello stesso Istituto 114 alunni, ma solo 99 hanno conseguito il diploma entro l'anno 2011. Si vuole verificare, al livello di significatività del 5%, se gli iscritti dell'anno 2006 hanno avuto un rendimento minore di quello degli iscritti ai corsi precedenti. Come si deve procedere?

RISOLUZIONE. Ancora delle considerazioni generali, quantunque riferite ad una situazione particolare.

Si assume come ipotesi nulla l'ipotesi che gli iscritti del 2006 abbiano avuto un rendimento non minore di quello del corso precedente. Questo vuol dire che la percentuale di coloro che hanno conseguito il diploma entro i 5 anni di scuola riferita agli iscritti del 2006 non è minore della percentuale riferita agli

iscritti dei corsi precedenti. Per cui, indicata con p la percentuale degli iscritti del 2006 che hanno conseguito il diploma entro il 2011 e con p_0 quella dei corsi precedenti, si hanno le due seguenti ipotesi:

- ipotesi nulla (H_0) : $p \geq p_0$;
- ipotesi alternativa (H_1) : $p < p_0$.

Con riferimento alla curva normale standardizzata, al livello di significatività del 5%, la regione dell'accettazione dell'ipotesi nulla è pari al 95% della regione sottostante la curva, ma situata questa volta tutta nella zona di destra di tale regione e precisamente a destra del valore $h=-1{,}65$ (Fig. 2).

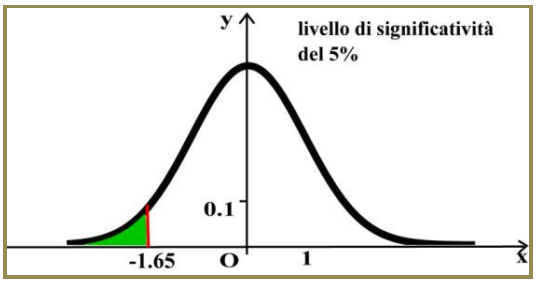

FIG. 2 – Test unilaterale sinistro

Il fatto che la regione del rifiuto dell'ipotesi nulla sia costituita dalla regione sottostante la curva normale situata a sinistra del valore $h=-1{,}65$ fa dire che si tratta di un ***test unilaterale sinistro***.

Per spiegare che il valore di h (evidentemente $h<0$) sia realmente questo, basta constatare che l'area $A(h)$ dell'accettazione è tale che: $A(h)=P[h \leq N \leq 0]+0{,}5$ e che deve essere $A(h)=0{,}95$. Di modo che si ha: $P[h \leq N \leq 0]=0{,}45$.

Ora, per ragioni di simmetria: $P[h \leq N \leq 0]=P[0 \leq N \leq -h]$, o anche, una volta posto $k=-h$: $P[h \leq N \leq 0]=P[0 \leq N \leq k]$. Per cui si deve trovare per quale valore di k risulta $P[0 \leq N \leq k]=0{,}45$. Per questo bisogna ricorrere alla tabella della distribuzione normale standardizzata (cfr.: unità 80, n. 80.1.3, Tab. 1). Siccome nella colonna delle probabilità di tale tabella non figura il valore 0,45 si opera una interpolazione lineare fra i due valori più vicini che comprendono 0,45 e che sono 0,44520 e 0,45543 cui corrispondono i valori 1,6 ed 1,7 di k. Si ha allora la seguente relazione:

$$\frac{k-1{,}6}{1{,}7-1{,}6}=\frac{0{,}45-0{,}44520}{0{,}45543-0{,}44520}$$

da cui segue $k \approx 1{,}65$. E quindi $h=-k=-1{,}65$.

Ritornando adesso al nostro problema, come nel caso precedente convertiamo la percentuale media campionaria nella variabile z standardizzata, tenendo presente che adesso, nella formula [4], al posto di μ, $\mu_{\overline{X}}$, $\sigma_{\overline{X}}$ ci vanno rispettivamente la percentuale p riferita al collettivo statistico, la percentuale media campionaria $\mu_{\overline{p}}$ e la deviazione standard da tale media $\sigma_{\overline{p}}$, per cui:

[5] $$z=\frac{\mu_{\overline{p}}-p}{\sigma_{\overline{p}}}\ .$$

Indicata allora con f la percentuale di iscritti nel 2006 che hanno conseguito il diploma entro il 2011, constatiamo che si ha:

$$p=p_0=0{,}925\ ,\quad \mu_{\overline{p}}=f=\frac{99}{114}\ ,\quad \sigma_{\overline{p}}=\sqrt{\frac{f(1-f)}{n}}=\sqrt{\frac{\frac{99}{114}\left(1-\frac{99}{114}\right)}{114}}\ .$$

Pertanto, mediante uno strumento di calcolo automatico:

$$z=\frac{\mu_{\overline{p}}-p}{\sigma_{\overline{p}}}=\frac{\frac{99}{114}-0{,}925}{\sqrt{\frac{\frac{99}{114}\left(1-\frac{99}{114}\right)}{114}}}\approx-1{,}78\,.$$

Siccome $z=-1{,}78<-1{,}65$ questo valore di z cade nella regione del rifiuto dell'ipotesi nulla. Si deve pertanto concludere che l'ipotesi nulla va rifiutata e va accettata invece l'ipotesi alternativa, vale a dire l'ipotesi che gli iscritti del 2006 hanno avuto un rendimento inferiore a quello dei corsi precedenti. Ovviamente al livello di significatività del 5%.

Se, infatti, tale livello fosse stato dell'1% (regione del rifiuto a sinistra del valore $k=-2{,}32$), essendo in tal caso $z>-2{,}32$ il valore di z sarebbe caduto nella regione dell'accettazione dell'ipotesi nulla e pertanto tale ipotesi sarebbe dovuta essere accettata.

81.4.3 Riassumiamo, e per alcuni aspetti integriamo, quanto abbiamo illustrato attraverso gli esercizi precedenti. Indichiamo per comodità con A, A^+, A^- le regioni del rifiuto dell'ipotesi nulla nel caso di un test rispettivamente *bilaterale*, *unilaterale destro* ed *unilaterale sinistro*.

Nel caso di livelli di significatività del 5%, dell'1% e del 10%, si hanno i risultati sintetizzati nella tabella sottostante (Tab. 3).

Regione rifiuto **Livello significatività**	**A**	**A^+**	**A^-**
5%	$z<-1{,}96$ o $z>1{,}96$	$z>1{,}65$	$z<-1{,}65$
1%	$z<-2{,}58$ o $z>2{,}58$	$z>2{,}32$	$z<-2{,}32$
10%	$z<-1{,}65$ o $z>1{,}65$	$z>1{,}28$	$z<-1{,}28$

TAB. 3

La determinazione dei valori numerici presenti in tale tabella, quando non sono già stati trovati, può essere effettuata con procedimenti simili a quello descritto nella risoluzione del precedente esercizio 2. Lasciamo a te questo eventuale compito, così come ti lasciamo quello di fornire le rappresentazioni grafiche delle regioni del rifiuto dell'ipotesi nulla nelle diverse situazioni, sempreché si tratti di fenomeni descrivibili con la distribuzione normale. Ti ricordiamo che noi abbiamo rappresentato i casi di un test bilaterale e di un test unilaterale sinistro, entrambi al livello di significatività del 5%.

81.4.4 Ribadiamo che quanto detto riguarda procedimenti in cui il campione in gioco ha numerosità almeno pari a 30. Questo per poter ricorrere alla distribuzione normale.

Naturalmente si potrebbero prendere in considerazione casi in cui l'esperimento idoneo a verificare una certa ipotesi sia condotto testando un campione di dimensione minore di 30. In tal caso bisogna far ricorso ad altre distribuzioni di probabilità. Vi faremo un cenno nel paragrafo n. 81.6.

Va detto poi che, oltre ai test relativi alla media ed alla proporzione di un collettivo statistico, si presentano solitamente anche altri test, come, ad esempio, quelli che riguardano la differenza fra due medie o fra due proporzioni. Di ciò, considerato il livello elementare della nostra trattazione, non ci occupiamo, rimandando a studi universitari in campo economico, in cui questi argomenti sono sviluppati ad un maggior livello di approfondimento.

81.5 LA STIMA [7]

81.5.1 Le stime dei parametri di una popolazione, fornite dalle formule [1] e [2] se l'indagine riguarda un aspetto quantitativo oppure dalle formule [1'] e [2'] se riguarda un aspetto qualitativo, sono stime espresse da un solo valore numerico. Ogni stima siffatta si chiama ***stima puntuale***.
La stima però può anche riferirsi ad un intervallo di valori e, contemporaneamente, alla probabilità che il parametro su cui s'indaga appartenga a quell'intervallo. Si parla allora di ***stima intervallo***. Ed è di questa che andiamo ad occuparci adesso.

Prenderemo in esame la situazione in cui il collettivo statistico su cui s'indaga abbia dimensione N "molto grande" ed anche la dimensione n del campione sia sufficientemente grande (in pratica $n \geq 30$), ma sia comunque $n < 0{,}05N$. In tal caso è possibile ricorrere alla distribuzione normale. Ancora una volta non ci soffermeremo sulle dimostrazioni.
Incominciamo ad indicare con g la probabilità P che la media μ di una popolazione statistica ha di cadere nell'intervallo $\left[\mu_{\overline{X}} - k\sigma_{\overline{X}}, \mu_{\overline{X}} + k\sigma_{\overline{X}}\right]$, dove $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ sono gli indici riferiti alla media campionaria e k è un parametro reale. In simboli:

$$P\left[\mu_{\overline{X}} - k\sigma_{\overline{X}} \leq \mu \leq \mu_{\overline{X}} + k\sigma_{\overline{X}}\right] = g\,.$$

Questo significa che, se si estraggono campioni casuali della popolazione, tutti della stessa dimensione, è uguale a g la percentuale di essi che presenta un intervallo $\left[\mu_{\overline{X}} - k\sigma_{\overline{X}}, \mu_{\overline{X}} + k\sigma_{\overline{X}}\right]$ nel quale è contenuta la media μ della popolazione.
La probabilità g è chiamata **livello di confidenza**. L'intervallo $\left[\mu_{\overline{X}} - k\sigma_{\overline{X}}, \mu_{\overline{X}} + k\sigma_{\overline{X}}\right]$ è chiamato **intervallo di confidenza**.
Analogamente, se invece di una media si considera la percentuale p della popolazione che presenta una determinata caratteristica, la probabilità P che tale percentuale cada nell'intervallo $\left[\mu_{\overline{p}} - k\sigma_{\overline{p}}, \mu_{\overline{p}} + k\sigma_{\overline{p}}\right]$ è la seguente:

$$P\left[\mu_{\overline{p}} - k\sigma_{\overline{p}} \leq p \leq \mu_{\overline{p}} + k\sigma_{\overline{p}}\right] = g\,.$$

Anche adesso la probabilità g è chiamata **livello di confidenza** e l'intervallo $\left[\mu_{\overline{p}} - k\sigma_{\overline{p}}, \mu_{\overline{p}} + k\sigma_{\overline{p}}\right]$ è chiamato **intervallo di confidenza**.
Il prodotto $k\sigma_{\overline{X}}$ (o eventualmente $k\sigma_{\overline{p}}$), dipendente ovviamente da k, è l'***errore*** ε che si commette nell'assumere $\overline{x}$ al posto di μ (ovvero f al posto di p). Dunque:

$$\overline{x} = \mu \pm k\sigma_{\overline{X}}\,, \quad f = p \pm k\sigma_{\overline{p}}\,.$$

La quantità $\sigma_{\overline{X}}$ (o eventualmente $\sigma_{\overline{p}}$), vale a dire la deviazione standard campionaria, ottenuta chiaramente per k=1, si chiama anche ***errore standard*** della media (o eventualmente della proporzione).
Vedremo fra breve com'è collegato l'errore ε alla probabilità g.

81.5.2 La probabilità g dipende dal parametro k. Per stabilire in che modo, incominciamo a risolvere il seguente problema.

- PROBLEMA. Determinare l'intervallo $[\mu - k\sigma, \mu + k\sigma]$ entro cui è compreso, con probabilità g assegnata, un valore della variabile aleatoria normale $X = N(\mu, \sigma^2)$.

[7] Questo paragrafo riguarda la 5ª classe sia di Liceo Scientifico, compresa l'opzione Scienze applicate, sia di Tecnici e Professionali.

In particolare determinare gli intervalli corrispondenti ai seguenti valori di g: 90%, 95%, 99%. Determinare, inoltre, quali probabilità g corrispondono ai valori di k: 1, 2, 3.

RISOLUZIONE. Si tratta di determinare per quale valore di k risulta: $P[\mu-\sigma \le X \le \mu+\sigma]=g$.

Incominciamo a constatare che, utilizzando una nota formula, si ha:

$$P[\mu-k\sigma \le X \le \mu+k\sigma]=P\left[\frac{(\mu-k\sigma)-\mu}{\sigma} \le N \le \frac{(\mu+k\sigma)-\mu}{\sigma}\right]=P[-k \le N \le k]=2\,P[0 \le N \le k].$$

Perciò:

$$P[0 \le N \le k] = \frac{g}{2}\,.$$

Pertanto, quando g=90%=0,90 bisogna trovare per quale valore di k risulta:

$$P[0 \le N \le k] = 0{,}45.$$

Ora, il valore 0,45 non compare nella seconda colonna della tabella delle probabilità normali standardizzate[8]. Cerchiamo allora i due valori consecutivi di k tra i quali esso è compreso. Sono 0,44520 cui corrisponde il valore k=1,6 e 0,45543 cui corrisponde k=1,7. Con un'interpolazione lineare fra questi valori si ottiene:

$$\frac{0{,}45-0{,}44520}{0{,}45543-0{,}44520}=\frac{k-1{,}6}{1{,}7-1{,}6}$$

da cui segue: $k \approx 1{,}65$.

In definitiva, se g = 90% allora l'intervallo cercato è $[\mu-1.65\sigma, \mu+1.65\sigma]$.

Si ragiona allo stesso modo negli altri casi, che lasciamo a te per esercizio.

Una tabella (Tab. 4) riassume la corrispondenza tra il valore k e l'intervallo $[\mu-k\sigma, \mu+k\sigma]$ entro cui è compreso un valore della variabile aleatoria normale $N(\mu, \sigma^2)$ con probabilità g e la probabilità medesima.

k	1	2	3	1,65	1,96	2,58
g	68,27%	95,45%	99,73%	90%	95%	99%

TAB. 4

Una riflessione, utile a mettere meglio a fuoco i concetti che questa tabella esprime.

Tenendo presente che g è il doppio dell'area A(k) sotto il grafico della funzione:

$$y = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

relativa all'intervallo [0,k], si ha che:

a) per k=1 e quindi g=68,27% l'area A(k) è il doppio di quella evidenziata in figura 3;
b) per k=2 e quindi g=95,45% l'area A(k) è il doppio di quella evidenziata in figura 4;
c) per k=3 e quindi g=99,73%, l'area A(k) è il doppio di quella evidenziata in figura 5;
d) per k=1,65 e quindi g=90% l'area A(k) è il doppio di quella evidenziata in figura 6;
e) per k=1,96 e quindi g=95% l'area A(k) è il doppio di quella evidenziata in figura 7;
f) per k=2,58 e quindi g=99% l'area A(k) è il doppio di quella evidenziata in figura 8.

[8] Vedere unità precedente.

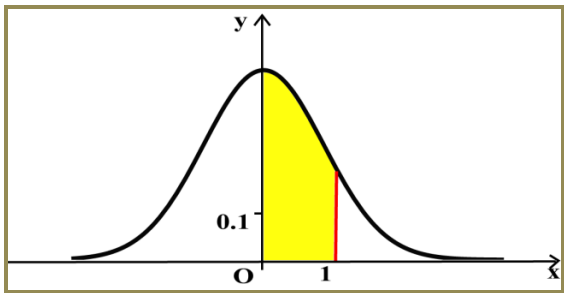

FIG. 3

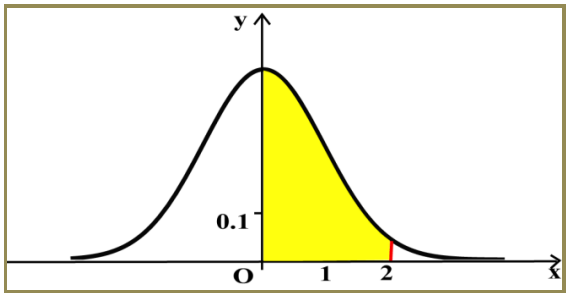

FIG. 4

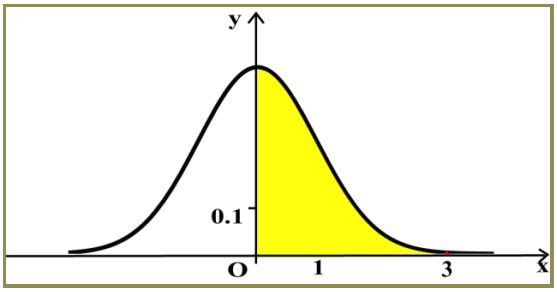

FIG. 5

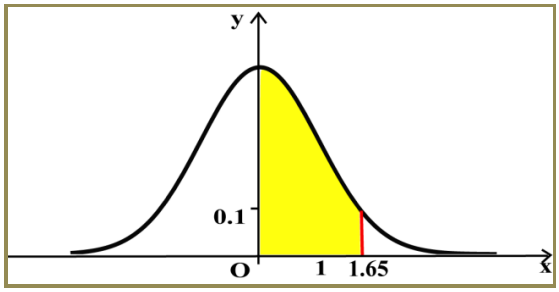

FIG. 6

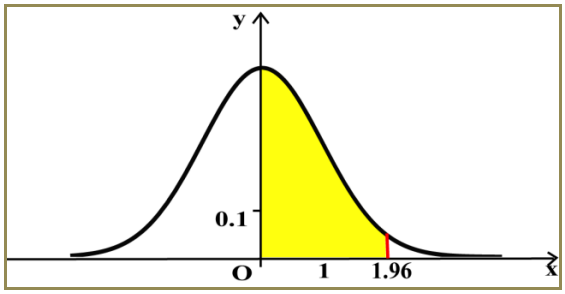

FIG. 7

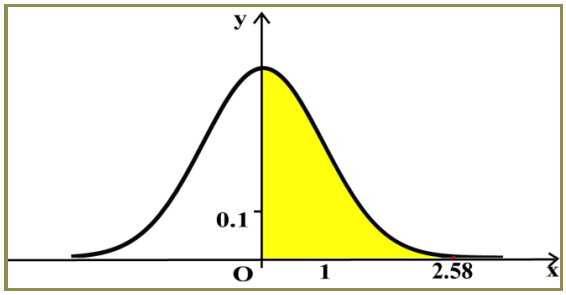

FIG. 8

81.5.3 Allora, adattando le considerazioni precedenti alla questione di statistica dalla quale siamo partiti e supponendo che l'indagine riguardi un aspetto di carattere quantitativo e perciò si pone l'attenzione sulla media campionaria $\overline{X}$, si ha che:

a) per k=1 si ha g=68,27% : la media della popolazione è compresa nell'intervallo $[\mu_{\overline{X}} - \sigma_{\overline{X}}\,,\ \mu_{\overline{X}} + \sigma_{\overline{X}}]$ con una probabilità del 68,27% e con un errore massimo pari a $\sigma_{\overline{X}}$;

b) per k=2 si ha g=95,45% : la media della popolazione è compresa nell'intervallo $[\mu_{\overline{X}} - 2\sigma_{\overline{X}}\,,\ \mu_{\overline{X}} + 2\sigma_{\overline{X}}]$ con una probabilità del 95,45% e con un errore massimo pari a $2\sigma_{\overline{X}}$;

c) per k=3 si ha g=99,73% : la media della popolazione è compresa nell'intervallo $[\mu_{\overline{X}} - 3\sigma_{\overline{X}}\,,\ \mu_{\overline{X}} + 3\sigma_{\overline{X}}]$ con una probabilità del 99,73% e con un errore massimo pari a $3\sigma_{\overline{X}}$;

d) per k=1,65 si ha g=90% : la media della popolazione è compresa nell'intervallo $[\mu_{\overline{X}} - 1{,}65\ \sigma_{\overline{X}}\,,\ \mu_{\overline{X}} + 1{,}65\ \sigma_{\overline{X}}]$ con probabilità del 90% e con un errore massimo pari a $1{,}65\sigma_{\overline{X}}$;

e) per k=1,96 si ha g=95% : la media della popolazione è compresa nell'intervallo $[\mu_{\overline{X}} - 1{,}96\ \sigma_{\overline{X}}\,,\ \mu_{\overline{X}} + 1{,}96\ \sigma_{\overline{X}}]$ con probabilità del 95% e con un errore massimo pari a $1{,}96\sigma_{\overline{X}}$;

f) per k=2,58 si ha g=99% : la media della popolazione è compresa nell'intervallo

$[\mu_{\overline{X}} - 2{,}58\ \sigma_{\overline{X}}\ ,\ \mu_{\overline{X}} + 2{,}58\ \sigma_{\overline{X}}]$ con probabilità del 99% e con un errore massimo pari a $2{,}58\sigma_{\overline{X}}$.

I livelli di confidenza del 68,27%, del 95,45% e del 99,73% sono solitamente usati nelle scienze sperimentali. Quelli del 90%, del 95% e del 99% lo sono nelle scienze sociali ed economiche.
Facciamo notare che, nel caso delle scienze sperimentali, quando si assume un livello di confidenza del 99,73%, in meno di 3 prove su 1.000 si commette un errore maggiore dell'errore massimo, che è $\varepsilon = 3\sigma_{\overline{X}}$, nell'accettare $\mu_{\overline{X}}$ come misura della grandezza in esame
Riassumiamo i risultati precedenti in una tabella (Tab. 5):

k	Livello di confidenza	Intervallo di confidenza	Errore massimo
1	68,27 %	$[\mu_{\overline{X}} - \sigma_{\overline{X}}\ ,\ \ \mu_{\overline{X}} + \sigma_{\overline{X}}]$	$\sigma_{\overline{X}}$
2	95,45 %	$[\mu_{\overline{X}} - 2\ \sigma_{\overline{X}}\ ,\ \ \mu_{\overline{X}} + 2\ \sigma_{\overline{X}}]$	$2\ \sigma_{\overline{X}}$
3	99,73 %	$[\mu_{\overline{X}} - 2\ \sigma_{\overline{X}}\ ,\ \ \mu_{\overline{X}} + 2\ \sigma_{\overline{X}}]$	$3\ \sigma_{\overline{X}}$
1,65	90 %	$[\mu_{\overline{X}} - 1{,}65\ \sigma_{\overline{X}}\ ,\ \ \mu_{\overline{X}} + 1{,}65\ \sigma_{\overline{X}}]$	$1{,}65\ \sigma_{\overline{X}}$
1,96	95 %	$[\mu_{\overline{X}} - 1{,}96\ \sigma_{\overline{X}}\ ,\ \ \mu_{\overline{X}} + 1{,}96\ \sigma_{\overline{X}}]$	$1{,}96\ \sigma_{\overline{X}}$
2,58	99 %	$[\mu_{\overline{X}} - 2{,}58\ \sigma_{\overline{X}}\ ,\ \ \mu_{\overline{X}} + 2{,}58\ \sigma_{\overline{X}}]$	$2{,}58\ \sigma_{\overline{X}}$

TAB. 5

I valori della tabella sono i medesimi se, invece di un aspetto quantitativo, l'indagine riguarda l'esistenza o meno di un determinato carattere. Solo che adesso al posto di $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ figurano rispettivamente $\mu_{\overline{p}}$ e $\sigma_{\overline{p}}$.

Si può notare come l'ampiezza dell'intervallo di confidenza aumenti al crescere del livello di confidenza prescelto, aumentando in questo modo anche la vaghezza dell'informazione. D'altronde se il livello di confidenza è troppo piccolo, può assumere un significato inconsistente. Per questo sono utilizzati per lo più il livello di confidenza del 95,45% nelle scienze sperimentali e quello del 95% in economia.

81.5.4 Presentiamo adesso alcune esemplificazioni di quanto siamo andati dicendo fin qui. Di tanto in tanto chiederemo anche la tua collaborazione.

- ESERCIZIO 1. In una città di 80.000 abitanti si vuole condurre una ricerca di mercato, al fine di confermare (o smentire) che un certo prodotto sembra trovare gradimento nel 60% della popolazione. Determinare la dimensione del campione su cui indagare affinché, al livello di confidenza g, l'indagine fornisca un risultato affetto da un errore massimo ε, sapendo che:

a) $g=95\%, \varepsilon=4\%$; b) $g=95\%, \varepsilon=3\%$; c) $g=99\%, \varepsilon=3\%$.

RISOLUZIONE. L'indagine riguarda evidentemente l'esistenza o meno di un determinato carattere. Nella fattispecie, il gradimento di un dato prodotto. Sappiamo che, indicata con $\sigma_{\overline{p}}$ la deviazione standard della media campionaria e chiamato ε l'errore massimo che si commette, in genere risulta: $k\sigma_{\overline{p}}=\varepsilon$, da cui segue: $\sigma_{\overline{p}}=\varepsilon/k$.
Inoltre, essendo abbastanza grande la dimensione della popolazione su cui s'indaga, sappiamo che si ha:

$$\sigma_{\overline{p}}=\sqrt{\frac{f(1-f)}{n}}\text{ , da cui segue: } n=\frac{f(1-f)}{\sigma_{\overline{p}}^2}\text{ e perciò: } n=\frac{f(1-f)\,k^2}{\varepsilon^2}$$

dove n è la dimensione del campione. D'altro canto, nel caso in esame è f=0,6 per cui:

$$n=\frac{0{,}6\times0{,}4\times k^2}{\varepsilon^2}=0{,}24\,\frac{k^2}{\varepsilon^2}\,.$$

Pertanto:

a) al livello di confidenza del 95%, cui corrisponde k=1,96, se si suppone ε=0,04 si ha:

$$n=0{,}24\,\frac{1{,}96^2}{0{,}04^2}=577\,;$$

b) analogamente, come puoi trovare da te: n=1025;

c) n=1776.

L'esempio mostra che la dimensione del campione aumenta con l'aumentare del livello di confidenza e col diminuire del margine di errore e questo vale in generale.
Ora, siccome analizzare un campione molto numeroso potrebbe comportare serie difficoltà (si immagini di dover provare i pezzi di ricambio prodotti da una ditta per stabilire la percentuale di quelli difettosi) o essere molto dispendioso (intervistare 1000 persone è certamente più economico che intervistarne 5000), spesso si preferisce accontentarsi di un livello di confidenza più basso nei risultati e di un margine di errore più alto. Dipende, in ogni caso, da ciò su cui s'indaga: se l'indagine riguarda un aspetto di cui si richiede una previsione la più accurata possibile, è indubbio che si debba ipotizzare un campione molto numeroso ed un margine di errore molto basso; se invece la previsione può essere anche grossolana, ci si può accontentare di un campione poco numeroso e di un margine di errore piuttosto alto.

♦ OSSERVAZIONE. Nell'esempio precedente abbiamo supposto di conoscere la percentuale della popolazione che presenta il carattere su cui s'indaga (p=60%). Non sempre però si ha questa fortuna. Ragion per cui, in tal caso, bisogna cercare una modalità di indagine alternativa.
Ebbene, se non si ha alcuna informazione su p, è opportuno supporre che il fenomeno abbia la massima variabilità possibile, per evitare di trarre conclusioni errate. Questo significa che bisogna supporre massimo il prodotto p(1−p). Il che, come puoi trovare facilmente, accade per p=q=0,5.

In definitiva, quando non si hanno informazioni sulla percentuale della popolazione che presenta il carattere su cui s'indaga, è opportuno supporre che il 50% della popolazione presenti il carattere su cui s'indaga e, ovviamente, il 50% non lo presenti. È, infatti, questa la situazione di massima variabilità.

• ESERCIZIO 2. Della lunghezza L di un regolo sono state compiute 30 misurazioni e si sono ottenuti i seguenti valori, espressi in cm:

25,8	25,6	26,0	25,6	26,1	25,7	25,4	25,8	25,3	25,9
26,2	25,7	26,1	25,4	25,8	25,4	25,9	26,0	25,7	25,6
26,0	25,8	25,5	25,7	25,5	26,2	25,8	25,7	25,6	26,0

Calcolare la lunghezza del regolo ad un livello di confidenza del:

a) 68,27%, b) 99,73%.

RISOLUZIONE. Qui l'indagine riguarda un aspetto quantitativo: la misura di un regolo. Si calcola la media $\overline{x}$ e la deviazione standard s di questo campione di misure:

$$\overline{x} \approx 25{,}77 \text{ cm}, \quad s \approx 0{,}04 \text{ cm}.$$

Siccome la dimensione del campione non è inferiore a 30, possiamo ammettere con buona approssimazione che le misure del regolo siano distribuite secondo la curva normale di Gauss, con media 25,77 cm e deviazione standard 0,04 cm.

Possiamo stimare, adesso, entro quale intervallo è compresa la media μ delle misure del regolo (e quindi la lunghezza stessa del regolo), ad un dato livello di confidenza. Per questo dobbiamo calcolare dapprima la media $\mu_{\overline{X}}$ e la deviazione standard $\sigma_{\overline{X}}$ della distribuzione della media campionaria.

Possiamo supporre che la dimensione n=30 dei campioni sia minore di 0,05N, dove N è il numero delle possibili misurazioni del regolo (teoricamente infinite). Per cui si ha:

$$\mu_{\overline{X}} = \overline{x} = 25{,}77 \text{ cm}, \quad \sigma_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{0{,}04}{\sqrt{30}} \approx 0{,}007 \text{ (cm)}.$$

Possiamo concludere che la lunghezza L del regolo è, in centimetri:

a) $L = 25{,}77 \pm 0{,}007$ cioè, approssimando: $25{,}76 \le L \le 25{,}78$ al livello di confidenza del 68,27%;

b) $L = 25{,}77 \pm 3 \times 0{,}007 = 25{,}77 \pm 0{,}021$ cioè $25{,}75 \le L \le 25{,}79$ al livello di confidenza del 99,73%.

- ESERCIZIO 3. Un campione casuale di 100 giovani di età compresa fra i 16 ed i 30 anni, frequentatori di discoteche, viene interpellato circa la spesa sostenuta mensilmente per recarsi in discoteca. Ne scaturisce una spesa media di € 78 per persona con una deviazione standard di € 18. Stabilire entro quali limiti si può stimare che vari la spesa sostenuta mensilmente da un giovane di quell'età, frequentatore di discoteche, per recarsi in discoteca, ad un livello di confidenza del:

a) 90%, b) 95%, c) 99%.

RISOLUZIONE. Di nuovo, l'indagine riguarda un aspetto quantitativo: la spesa sostenuta per frequentare le discoteche. Cominciamo a calcolare la media e la deviazione standard delle medie campionarie:

$$\mu_{\overline{X}} = \overline{x} = 78 \text{ (€)}, \quad \sigma_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{18}{\sqrt{100}} \approx 1{,}8 \text{ (€)}.$$

Osservato che la spesa S può essere stimata inclusa nell'intervallo [78–1.8k, 78+1.8k] ad un livello di confidenza g, si ha:

a) per g=90%, nel qual caso k=1,65 e perciò 1,8k=2,97 (€), risulta: S=78±2,97 (€), vale a dire che la spesa S, espressa in euro, è compresa entro i limiti: $75{,}03 \le S \le 80{,}97$;

b) per g=95%, nel qual caso k=1,96 e perciò 1,8k=3,52 (€), risulta: S=78±3,53 (€), vale a dire che la spesa S, espressa in euro, è compresa entro i limiti: $74{,}47 \le S \le 81{,}53$;

c) per g=99%, nel qual caso k=2,58 e perciò 1,8k=4,64 (€), risulta: S=78±4,64 (€), vale a dire che la spesa S, espressa in euro, è compresa entro i limiti: $73{,}36 \le S \le 82{,}64$.

- ESERCIZIO 4. In seguito ad un'indagine svolta intervistando un campione casuale di 500 elettori di una regione, si scopre che il 33,2% degli interpellati dichiara che alle prossime elezioni voterà per un determinato partito. Stimare entro quali limiti è compresa la percentuale di voti che prenderà quel partito, ad un livello di confidenza del:

a) 90%, b) 95%, c) 99%.

RISOLUZIONE. Adesso l'indagine riguarda la presenza o meno di un determinato carattere: la dichiarazione di voto per un dato partito. La percentuale P di voti che il partito in questione prenderà può essere stimata inclusa nell'intervallo $P[\mu_{\overline{p}} - k\sigma_{\overline{p}}, \ \mu_{\overline{p}} + k\sigma_{\overline{p}}]$, dove:

$$\mu_{\overline{p}} = f = 0{,}332, \quad \sigma_{\overline{p}} = \sqrt{\frac{f(1-f)}{n}} = \sqrt{\frac{0{,}332(1-0{,}332)}{500}} \approx 0{,}02719$$

mentre k dipende dal livello di confidenza g che viene stabilito. Si ha che:

a) per g=90%, nel qual caso k=1,65 e perciò $k\sigma_{\overline{p}} \approx 1{,}65 \times 0{,}02719 \approx 0{,}035$, risulta P = 0,332±0,035, vale a dire che si ha: 29,7% ≤ P ≤ 36,7%;

b) per g=95%, nel qual caso k=1,96 e perciò $k\sigma_{\overline{p}} \approx 1{,}96 \times 0{,}02719 \approx 0{,}041$, risulta P = 0,332±0,041, vale a dire che si ha: 29,1% ≤ P ≤ 37,3%;

c) per g=99%, nel qual caso k=2,58 e perciò $k\sigma_{\overline{p}} \approx 2{,}58 \times 0{,}02719 \approx 0{,}054$, risulta P = 0,332±0,054, vale a dire che si ha: 27,8% ≤ P ≤ 38,6%.

81.6 STIME PER PICCOLI CAMPIONI [9]

81.6.1 Quando la dimensione del campione è minore di 30, il collettivo statistico dal quale si suppone estratto il campione non si può automaticamente considerare distribuito normalmente, come invece accade quando tale dimensione non è inferiore a 30. Ragion per cui, in tal caso, non si può fare riferimento alla distribuzione di Gauss. Se però ci sono elementi per supporre che comunque il collettivo sia distribuito normalmente, allora si può fare ricorso ad una distribuzione particolare, detta **distribuzione t di Student**, per ottenere gli intervalli di confidenza per la media di una grandezza.
La densità di probabilità p(x) della distribuzione di Student di parametro n è fornita dalla seguente relazione:

$$\mathbf{p(x) = C(n) \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}}$$

dove C(n) è una funzione di n. I valori di C(n), calcolati per n (ovviamente intero) compreso fra 4 e 29 inclusi, sono forniti dalla tabella sottostante (Tab. 6).

n	C(n)	**n**	C(n)	**n**	C(n)	**n**	C(n)
4	0,37500	**11**	0,38998	**18**	0,39344	**25**	0,39497
5	0,37960	**12**	0,39072	**19**	0,39372	**26**	0,39512
6	0,38273	**13**	0,39135	**20**	0,39398	**27**	0,39526
7	0,38449	**14**	0,39188	**21**	0,39422	**28**	0,39539
8	0,38669	**15**	0,39235	**22**	0,39443	**29**	0,39551
9	0,38803	**16**	0,39276	**23**	0,39463		
10	0,38910	**17**	0,39312	**24**	0.39480		

TAB. 6

Detto per curiosità ed a beneficio di chi avesse voglia di verificare i risultati della precedente tabella, naturalmente utilizzando un programma di calcolo automatico, segnaliamo che si ha:

$$C(n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)},$$

dove Γ(α) è la funzione, detta *funzione Gamma di Eulero*, così definita:

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx\,.$$

[9] Questo paragrafo riguarda soltanto l'indirizzo "Sistema Moda" dell'Istituto Tecnico, settore Tecnologico.

Bisogna precisare, per evitare ogni equivoco, che n non è la dimensione del campione, bensì quello che viene definito il suo *grado di libertà*. La dimensione D del campione è tale che n=D−1.
Nel caso particolare in cui n=10 (e quindi dimensione D=11), per cui risulta C(10)=0,38910, la distribuzione t (di parametro 10) ha la seguente densità:

$$p(x) = 0{,}38910 \cdot \left(1 + \frac{x^2}{10}\right)^{-5{,}5}.$$

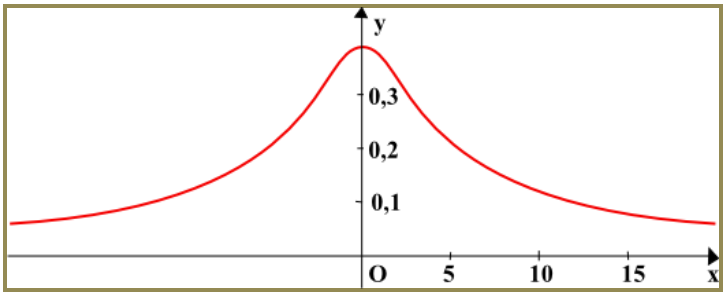

FIG. 9

La funzione è rappresentata in figura 9 ed appare molto simile alla curva di Gauss. In realtà le differenze sostanziali si presentano nelle "code": la curva di Gauss è più schiacciata sull'asse delle ascisse rispetto a quella di Student, per cui l'area sotto il grafico di quest'ultima è, in corrispondenza delle code, maggiore di quella che c'è sotto il grafico della curva di Gauss.

Student è lo pseudonimo sotto il quale il matematico e statistico inglese **William Sealy Gosset** (1876-1937) pubblicò un suo lavoro relativo alle caratteristiche della distribuzione t. La denominazione "distribuzione di Student" sarebbe venuta poco tempo dopo per merito di **Fisher**.

81.6.2 Anche adesso, come nel caso della distribuzione di Gauss, ma con stime più grossolane, si può calcolare la probabilità g che la media μ del collettivo statistico ha di cadere nell'intervallo $[\mu_{\overline{X}} - k\sigma_{\overline{X}}, \mu_{\overline{X}} + k\sigma_{\overline{X}}]$. Si ha ovviamente:

$$P[\mu_{\overline{X}} - k\sigma_{\overline{X}} \le \mu \le \mu_{\overline{X}} + k\sigma_{\overline{X}}] = g,$$

dove k varia a seconda non solo del variare di g ma anche del variare di n, secondo la seguente legge:

[6] $$\mathbf{C(n)\int_0^k \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} dx = \frac{g}{2}}.$$

Una tabella (Tab. 7) evidenzia il legame determinato dalla [6] per n che varia da 4 a 13 (e quindi per D che varia da 5 a 14) e per i seguenti valori di g: 90%, 95%, 99%. Altri eventuali valori di k puoi determinarli da solo, utilizzando naturalmente un idoneo software matematico.

g n=D−1	0,90	0,95	0,99
4	2,132	2,776	4,604
5	2,015	2,571	4,032
6	1,943	2,447	3,707
7	1,895	2,365	3,499
8	1,860	2,306	3,355

9	1,833	2,262	3,250
10	1,812	2,228	3,169
11	1,796	2,201	3,106
12	1,782	2,179	3,055
13	1,771	2,160	3,012

TAB. 7

- ESERCIZIO. Misurando la lunghezza di un regolo, si ottengono le seguenti 11 misure, espresse in centimetri:

77,45 77,11 77,93 77,88 77,92 77,04 77,80 77,59 77,58 77,12 77,95.

Qual è la lunghezza del regolo ad un livello di confidenza del 90% ?

RISOLUZIONE. Le 11 misure ottenute possono concepirsi come un campione di dimensione 11 delle (infinite) possibili misure del regolo. Ragion per cui, mentre la distribuzione di tutte le possibili misure si può ritenere normale, non si può far riferimento alla distribuzione normale per quanto riguarda il campione ($D<30$). Si può ricorrere tuttavia alla distribuzione di Student.

Ebbene, tenendo presente la tabella 5 (per $n=D-1=10$), possiamo affermare che, ad un livello di confidenza del 90%, la lunghezza del regolo è compresa fra i valori $\mu_{\bar{x}}-1{,}812\,\sigma_{\bar{x}}$ e $\mu_{\bar{x}}+1{,}812\,\sigma_{\bar{x}}$.

Si tratta allora di conoscere i due indici statistici $\mu_{\bar{x}}$ e $\sigma_{\bar{x}}$. Si trova, ovviamente con l'uso di uno strumento di calcolo automatico: $\mu_{\bar{x}}\approx 77{,}579$; $\sigma_{\bar{x}}\approx 0{,}354$. Di conseguenza, essendo:

$$\mu_{\bar{x}}-1{,}812\,\sigma_{\bar{x}}\approx 76{,}94 \quad \text{e} \quad \mu_{\bar{x}}+1{,}812\,\sigma_{\bar{x}}\approx 78{,}22$$

la lunghezza L del regolo, ad un livello di confidenza del 90%, è tale che:

$$76{,}94 \le L \le 78{,}22$$

Oppure, scritto in forma equivalente, dopo aver calcolato che $1{,}812\sigma_{\bar{x}}\approx 0{,}64$:

$$L = 77{,}58 \pm 0{,}64.$$

81.6.3 Ti proponiamo alcuni esercizi su questo argomento. Non ne troverai altri nella sezione "verifiche" né ci saranno domande in proposito nella "breve sintesi per domande e risposte".

a. Mediante l'uso di un idoneo software matematico, completa la seguente tabella, riferita ad un campione di dimensione 11, tenendo presente che i significati di k, g sono quegli stessi specificati sopra.

k	1	2	3			
g				0,6827	0,9545	0,9973

[**R.** $k=1\to g\approx 0{,}6591$; $k=2\to g\approx 0{,}9266$; $k=3\to g\approx 0{,}9866$; $g=0{,}6827\to k\approx 1{,}05$; $g=0{,}9545\to k\approx 2{,}28$; $g=0{,}9973\to k\approx 3{,}96$]

b. Nel laboratorio di fisica hai eseguito per 11 volte un esperimento nel tentativo di trovare una stima attendibile di una grandezza G. Hai quindi calcolato la media aritmetica $\mu_{\bar{x}}$ delle misure ottenute e la deviazione standard $\sigma_{\bar{x}}$. È corretto affermare che la misura di G è $\mu_{\bar{x}}\pm\sigma_{\bar{x}}$, ad un livello di confidenza del 68,27% ?

[**R.** Non è corretto. Al livello di confidenza del 68,27% la misura di G è $\mu_{\bar{x}}\pm 1{,}05\sigma_{\bar{x}}$; invece la misura di G è $\mu_{\bar{x}}\pm\sigma_{\bar{x}}$ ad un livello di confidenza del 65,91%.]

c. Nel laboratorio di fisica hai eseguito per 6 volte un esperimento nel tentativo di trovare una stima di una grandezza G. Qual è l'errore massimo al livello di confidenza del 68,27% ?

[**R.** $1{,}11\sigma_{\bar{x}}$]

d. Un campione casuale di 21 lampade alogene ha una durata media di 2000 ore di funzionamento con

una deviazione standard di 70 ore. La durata delle lampade della partita da cui è estratto il campione ha una distribuzione normale. Calcolare quale stima si può dare della durata delle lampade della partita al livello di confidenza del:

a) 90% ; b) 95% ; c) 99% .

[**R.** g=90%→k≈1,72; g=95%→k≈2,08; g=99%→k=2,84%]

e. Un campione casuale di 26 ragazzi di 18 anni ha un'altezza media di 174,3 cm con una deviazione standard di 3,2 cm. Le altezze del collettivo costituito da tutti i ragazzi di 18 anni sono distribuite normalmente. Calcolare quale stima si può dare dell'altezza dei ragazzi del collettivo in questione al livello di confidenza del:

a) 90% ; b) 95% ; c) 99% .

[**R.** g=90%→H=174,3±5,4; g=95%→H=174,3±6,6; g=99%→H=174,3±8,9]

VERIFICHE

Campionamento (nn. 1-4)

1. Un certo partito politico si propone di valutare quanti, fra i 45.700 elettori di una data città, lo voteranno alle prossime elezioni, sapendo che nelle elezioni precedenti i votanti a favore erano stati il 15,8%. Trovare quale deve essere la dimensione minima del campione da scegliere, se i risultati dell'indagine possono essere affetti da un errore massimo dell'8% e se su di essi si può fare affidamento al livello di confidenza del:

a) 90%; b) 95%; c) 99%.

[**R.** a) 57; b) 80; c) 139]

2. Un partito politico di nuova formazione vuole valutare quanti, fra i 15.800 elettori di una certa città, lo voteranno alle prossime elezioni. Determinare la dimensione minima del campione su cui condurre l'indagine, ammesso che i risultati possano essere affetti da un errore massimo del 10% e posto che su di essi si possa fare affidamento al livello di confidenza del:

a) 90%; b) 95%; c) 99%.

[**R.** Tenere presente che non si hanno informazioni sulla percentuale della popolazione che ... a) 69; b) 97; c) 167]

3. Una fabbrica di lampadine elettriche si propone di controllare se ve ne sono di difettose fra quelle che ha prodotto in un certo periodo. Per questo sceglie a caso, fra le lampadine prodotte, un campione rappresentativo. Trovarne la dimensione minima, posto che, sulla base delle esperienze passate, la fabbrica può stimare che, al livello di confidenza del 95%, la percentuale delle lampadine difettose sia pari all'1,5% dei pezzi prodotti, con un margine di errore:

a) del 4%; b) del 3%; c) del 2%.

[**R.** a) 36; b) 64; c) 142]

4. Si vuole indagare su una popolazione di 3000 unità riguardo all'esistenza di un certo carattere che in passato si è manifestato 2 volte su 5 nei soggetti che compongono la popolazione.
Determinare la dimensione minima del campione su cui bisogna condurre l'indagine, ammesso che i risultati possano essere affetti da un errore massimo del 4% e che su di essi si possa fare affidamento al livello di confidenza del:

a) 90%; b) 95%; c) 99%.

[**R.** a) 409; b) 577; c) 999]

Controllo di qualità (nn. 5-9)

5. La ditta che rifornisce di lattine di aranciata il minibar della scuola assicura che ogni lattina contiene almeno 330 cm^3 di bevanda, con una tolleranza del 2% sul numero delle lattine conformi. In seguito alle lamentele degli studenti, il consiglio d'istituto decide di effettuare un controllo. Fa esaminare perciò il contenuto di un campione casuale di 30 lattine e si trova che la percentuale di quelle conformi (almeno 330 cm^3 di bevanda) è dell'86,7%. A quale conclusione si giunge?

6. Un'azienda produce bulloni (vite più madrevite). La filettatura esterna delle viti deve avere un diametro compreso fra 11,701 mm e 11,966 mm. Per valutare se ci sono scarti l'azienda decide di controllare un campione casuale di 30 viti e trova che la media aritmetica delle filettature dei loro diametri esterni è di 11,853 mm con una deviazione standard di 0,096 mm. A quale conclusione giunge l'azienda? [**R.** Tolleranza = 0,1325 mm; quindi …]

7. Ciascuna delle confezioni speciali di pasta dovrebbe avere un peso di 3 kg con una deviazione standard dello 0,50% sul peso della stessa. Il supermercato, che ha acquistato un lotto di 2000 confezioni, decide di effettuare un controllo. Per questo monitora 30 confezioni di pasta, scelte a caso, e trova i risultati sintetizzati nella tabella sottostante. Cosa conclude il supermercato?

peso (kg)	2,970	2,988	2,995	2,999	3,004
N° pacchi	2	4	12	9	3

8. Un'azienda chimica produce mangimi per animali. Ogni confezione che immette sul mercato contiene un miscuglio di due mangimi che devono stare in un rapporto compreso fra 0,76 e 0,84. Per valutare se ci sono confezioni non conformi, l'azienda analizza un campione casuale di 30 confezioni e trova i risultati sintetizzati nella tabella sottostante. A quale conclusione giunge l'azienda?

rapporto	0,76	0,78	0,79	0,82	0,84
N° pacchi	3	5	10	8	4

9. Un'azienda farmaceutica produce un vaccino antiallergico, che dovrebbe risultare efficace sull'80% dei soggetti ai quali viene somministrato, con una tolleranza del 5% su tale numero. Per valutare se le sue aspettative sono giuste, monitora per 5 anni un campione di 200 persone che soffrono di quel tipo di allergia e trova che il vaccino ha avuto effetto su 150 di essi. A quale conclusione giunge l'azienda?

Test delle ipotesi (nn. 10-24)

10. Voglio controllare, al livello di significatività del 5%, l'ipotesi che la probabilità che esca "Testa", nel lancio di una moneta, sia 1/2. Per questo effettuo 30 lanci della moneta ed ottengo 20 volte "Testa" e 10 volte "Croce". Cosa devo concludere riguardo all'ipotesi formulata? Cosa al livello di significatività del 10%? [**R.** Test bilaterale: $z \approx 1{,}93$]

11. Con riferimento allo stesso esercizio precedente, cambiano i risultati se invece di 30 lanci ne effettuo 60, ma ottengo la stessa percentuale di "Testa" e la stessa di "Croce"? [**R.** $z \approx 2{,}73$]

12. Voglio controllare, al livello di significatività dell'1%, l'ipotesi che un dado sia truccato. Per questo effettuo 50 lanci del dado ed esce per 4 volte la faccia "1". Cosa devo concludere riguardo all'ipotesi formulata? Cosa al livello di significatività del 5%? [**R.** Test bilaterale: $z \approx -2{,}25$]

13. Un'azienda produttrice di latte in scatola, in seguito alle lamentele dei consumatori, ha deciso di effettuare un test, al livello di significatività dell'1%, per verificare che la quantità di latte contenuto nelle scatole non è inferiore al litro dichiarato sull'etichetta. Per questo esamina 30 scatole di latte

scelte casualmente e, fatte le debite misurazioni, trova che la quantità media di latte per scatola è 0,989 l con una deviazione standard di 0,045 l. Cosa deve concludere l'azienda riguardo all'ipotesi formulata? Cosa al livello di significatività del 10%? [**R.** Test unilaterale sinistro: $z \approx -1{,}33$]

14. Con riferimento allo stesso esercizio precedente, cambiano i risultati se invece di 30 scatole di latte l'azienda ne controlla 50, ottenendo la stessa media e la stessa deviazione standard? [**R.** $z \approx -1{,}72$]

15. Un'indagine, effettuata su 40 confezioni, mostra un peso medio del loro contenuto di 2,4 kg con una deviazione standard di 300 g. Cosa devo concludere, al livello di significatività del 5%, riguardo all'ipotesi che il peso delle confezioni dovrebbe essere di 2,5 kg? Cosa al livello di significatività dell'1%? [**R.** Test bilaterale: $z \approx -2{,}10$]

16. Il quoziente d'intelligenza (QI) degli impiegati di una grossa ditta specializzata è stato in passato mediamente uguale a 125. Dopo l'ultimo turnover sono entrati molti nuovi impiegati e la ditta vuole sottoporre a test, ad un livello di significatività del 5%, l'ipotesi che il valore medio del QI degli impiegati non è cambiato. Sceglie allora un campione casuale di 30 impiegati e trova che il valore medio del QI è 117 con una deviazione standard di 18. Cosa deve concludere la ditta riguardo all'ipotesi formulata? Cosa al livello di significatività dell'1%? [**R.** Test bilaterale: $z \approx -2{,}43$]

17. L'altezza media degli abitanti di una comunità, di età non inferiore ai 18 anni, aveva fatto registrare il valore di 170 cm nell'ultimo rilevamento. Essendo trascorso molto tempo da tale rilevamento, il responsabile della comunità vuole sottoporre a test, al livello di significatività del 5%, l'ipotesi che l'altezza media degli abitanti non sia diminuita. Per questo sceglie un campione casuale di 40 soggetti e, fatte le dovute misurazioni, trova che la loro altezza media è 175 cm con una deviazione standard di 14 cm. Cosa deve concludere? Cosa al livello di significatività dell'1%? [**R.** Test unilaterale sinistro: $z \approx 2{,}25$]

18. Nell'ultima indagine effettuata nel comune era emerso che il reddito annuo pro-capite era stato mediamente di € 16430. Il sindaco vuole sottoporre a test, al livello di significatività del 5%, l'ipotesi che tale reddito sia diminuito a causa della recente crisi. Per questo estrae a sorte 80 soggetti e, dopo i dovuti rilevamenti e le necessarie misurazioni, trova che il loro reddito pro-capite è mediamente di € 16750 con una deviazione standard di € 1410. Cosa conclude? Cosa al livello di significatività dell'1%? [**R.** Test unilaterale destro: $z \approx 2{,}03$]

19. Con riferimento allo stesso esercizio precedente, cambiano i risultati se invece di 80 soggetti il sindaco ne esamina 50, ottenendo la stessa media e la stessa deviazione standard? [**R.** $z \approx 1{,}60$]

20. Un'azienda di trasporti pubblici, in seguito alle lamentele dei clienti, decide di effettuare un test, al livello di significatività dell'1%, per verificare che la corsa contestata dura non più di 75 minuti. Per questo, dopo aver registrato le durate di 30 corse, trova che la loro media è di 79 minuti con una deviazione standard di 8 minuti. Cosa conclude? [**R.** Test unilaterale destro: $z \approx 2{,}73$]

21. Un'azienda profumiera lancia un nuovo prodotto perché convinta che almeno il 30% della popolazione lo trovi apprezzabile? Ma, per maggior sicurezza, effettua un test, al livello di significatività del 10%, per valutare se la sua ipotesi è accettabile? Per questo conduce un sondaggio, intervistando 400 potenziali acquirenti e trova che solo 105 dichiarano il loro apprezzamento mentre 295 si dicono insoddisfatti. Cosa conclude l'azienda? [**R.** Test unilaterale sinistro. $z \approx -1{,}47$]

22. Un'azienda produce funi d'acciaio con un carico di rottura dichiarato di $1570\ \mathrm{N/mm^2}$ (Newton per millimetro quadrato) ed una tolleranza del 4%. In seguito alle proteste di alcuni clienti che lamentano carichi di rottura inferiori a quello dichiarato, l'azienda decide di effettuare un test di verifica, al livello di significatività dell'1%, per verificare che i carichi di rottura delle funi prodotte non sono inferiori a quelli dichiarati. Per questo estrae un campione casuale di 30 funi e di ciascuna di esse misura il carico di rottura, ottenendo i risultati registrati nella tabella 1 riportata nell'esercizio propo-

sto in chiusura del precedente paragrafo 80.3.2. Qual è la conclusione?

[**R.** ipotesi nulla (H_0): $\mu \geq 1507{,}2$; ipotesi alternativa: (H_1): $\mu < 1507{,}2$; ...]

23. Un'azienda conserviera produce pomodori pelati in scatola. Su ogni barattolo è dichiarato il peso netto di 400 g ed un peso sgocciolato di 240 g. In seguito alle proteste di alcuni clienti che lamentano un peso della parte liquida superiore a 160 g, l'azienda decide di effettuare un test di verifica, al livello di significatività dell'1%, per verificare che il peso della parte liquida non supera 160 g. Per questo estrae un campione casuale di 30 barattoli e per ciascuno di essi misura la parte liquida, ottenendo i risultati (pesi espressi in grammi) registrati nella tabella sottostante. Cosa conclude?

Peso (g)	136	137	138	139	140	141	142	143	144
Numero barattoli	1	3	3	4	7	6	3	2	1

24. L'altezza media degli studenti dell'Istituto "Einstein" era di 168 cm l'anno passato. Dopo l'uscita degli studenti dell'ultima classe e l'ingresso dei nuovi della prima classe, l'altezza media sarà variata o è rimasta immutata? Per scoprirlo si effettua un test, al livello di significatività del 5%. Per questo si sceglie un campione casuale di 30 studenti e si misurano le loro altezze, la cui media risulta essere di 172 cm con una deviazione standard di 12 cm. Cosa si conclude riguardo all'ipotesi che le altezze non sono cambiate?

La stima (nn. 25-34)

25. Un partito politico si propone di condurre un'indagine per stimare quale sarà la percentuale di voti favorevoli ad esso nelle prossime elezioni, sapendo che nelle elezioni precedenti quella percentuale era stata del 24,5%. Per questo decide di scegliere un campione rappresentativo degli elettori.

a) Determinare la dimensione minima n del campione, ammesso che i risultati possano essere affetti da un errore massimo del 6% e che su di essi si possa fare affidamento al livello di confidenza del 90%.

b) L'indagine, condotta proprio su un campione rappresentativo della dimensione n trovata, rileva che voterà per quel partito il 20,4% degli elettori. Calcolare entro quali limiti è compresa la percentuale di tutti gli elettori che si stima voteranno quel partito, sempre al livello di confidenza del 90%.

[**R.** a) 140; b) $14{,}8\% \leq P \leq 26{,}0\%$]

26. Un'azienda vuole lanciare un certo prodotto sul mercato senza avere alcuna idea sulle preferenze degli eventuali acquirenti. Per questo decide di condurre un'indagine preventiva su un campione scelto a caso.

a) Determinare la dimensione minima n del campione, supponendo che i risultati dell'indagine possano essere affetti da un errore massimo del 10% e che su di essi si possa fare affidamento al livello di confidenza del 90%.

b) L'indagine, condotta proprio su un campione casuale di n soggetti, rileva che la percentuale di quelli favorevoli al prodotto è del 12%. Calcolare una stima dell'intervallo in cui è compresa la percentuale di tutte le persone favorevoli al prodotto, sempre al livello di confidenza del 90%.

[**R.** a) 69; b) $5{,}5\% \leq P \leq 18{,}5\%$]

27. Sui 38.500 iscritti ad un esame di concorso, si vuole valutare quanti sono in possesso di una laurea. Le ultime esperienze mostrano che, su 100 concorrenti, 9 sono laureati, con un margine di errore del 10%.

Determinare la dimensione del campione su cui condurre l'indagine, ammesso che i risultati si possano accettare al livello di confidenza del: a) 95%; b) 99%. [**R.** a) 32; b) 55]

28. Sui 41.300 iscritti ad un esame di concorso, si vuole valutare quanti sono in possesso di una laurea. Si deve scegliere un campione su cui condurre l'indagine, senza poter fare affidamento su passate esperienze. Si richiede che i risultati siano affidabili al livello di confidenza del 90% e siano affetti da un errore massimo del: a) 15%; b) 10%; c) 5%.
Trovare la dimensione del campione. [**R.** a) 31; b) 69; c) 273]

29. Una fabbrica di pezzi di ricambio si propone di controllare se ve ne sono di difettosi nell'ultimo stock. Per questo sceglie a caso un campione dei pezzi prodotti. Trovarne la dimensione posto che, sulla base delle passate esperienze, la fabbrica stima che, al livello di confidenza del 90% la percentuale di pezzi difettosi sia pari al 2% dei pezzi prodotti, con un margine di errore del: a) 3%; b) 2%.
[**R.** a) 60; b) 134]

30. Un certo partito politico vuole stimare quanti, fra i 48.000 elettori di una data città, lo voteranno, sapendo che nelle precedenti elezioni i votanti a favore erano stati in proporzione di 14 su 100. Trovare quale deve essere la dimensione minima del campione da scegliere, se i risultati possono essere affetti da un errore massimo del 5% e se su di essi si vuol fare affidamento al livello di confidenza del: a) 90%; b) 95%. [**R.** a) 132; b) 186]

31. Un partito politico di nuova formazione vuole stimare quanti, fra i 134.000 elettori di una certa città, lo voteranno alle prossime elezioni. Naturalmente il partito, che è nuovo, non può affidarsi a risultati precedenti. Trovare quale dimensione minima deve avere il campione su cui condurre l'indagine, ammesso che i risultati possano essere affetti da un errore del 10% e su di essi si possa fare affidamento al livello di confidenza del: a) 90%; b) 95%. [**R.** a) 69; b) 97]

32. Un'azienda vuole lanciare un certo prodotto e per questo intende condurre preventivamente un'indagine di mercato. Trovare la dimensione minima del campione su cui condurre l'indagine medesima, posto che:

a) al livello di confidenza del 95%, si stima che la percentuale di individui favorevoli al prodotto sia del 42% con un margine di errore del 3%;
b) al livello di confidenza del 90%, si stima che la percentuale di individui favorevoli al prodotto sia del 42% con un margine di errore del 3%;
c) al livello di confidenza del 90%, si stima che la percentuale di individui favorevoli al prodotto sia del 42% con un margine di errore del 2%;
d) non si hanno elementi sulla percentuale di individui favorevoli al prodotto, ma si è disposti ad accettare, al livello di confidenza del 90%, risultati affetti da un errore massimo del 2%.

33. L'ufficio di statistica di una città vuole valutare quale percentuale di cittadini presenta un determinato carattere, del quale ha avuto qualche sentore ma senza disporre di dati precedenti. Per questo decide di condurre un'indagine su un campione rappresentativo dei cittadini.

a) Determinare la dimensione minima n del campione, ammesso che i risultati possano essere affetti da un errore massimo del 5% e che su di essi si possa fare affidamento al livello di confidenza del 90%.
b) L'indagine, condotta proprio su un campione casuale di n cittadini, rileva che il 65.6% di essi presenta il carattere su cui s'indaga. Stimare l'intervallo nel quale è compresa la percentuale di tutti i cittadini che presentano quel carattere, sempre al livello di confidenza del 90%.

[**R.** a) 273; b) $60{,}9\% \le P \le 70{,}4\%$]

34. Nella tabella sottostante sono elencati i punteggi riportati da 30 studenti del Liceo Galilei, scelti a caso tra coloro che hanno superato l'esame di Stato nell'ultimo anno scolastico (e che sono stati complessivamente 297):

60	75	96	62	68	73	82	98	100	74
72	65	84	75	66	60	72	82	84	68
60	60	64	100	70	87	94	96	77	81

Determinare un intervallo di confidenza al 95% per il punteggio medio conseguito dagli studenti del Liceo Galilei che hanno superato l'esame di Stato. [**R**. $71{,}6 < \mu < 79{,}8$]

UNA BREVE SINTESI PER DOMANDE E RISPOSTE

DOMANDE.

1. Un ragionamento si dice induttivo se è basato sul principio d'induzione. È vero o falso?
2. Che differenza c'è fra parametro e statistica?
3. È vero che un campione degli studenti della tua scuola è rappresentativo della stessa se è formato dagli studenti che hanno un rendimento scolastico medio?
4. È vero che per dimensione di un campione statistico s'intende la capacità del campione di rappresentare adeguatamente il collettivo da cui il campione è ottenuto?
5. Che cos'è la distribuzione della media campionaria?
6. Si supponga che la popolazione su cui si indaga abbia dimensione N, media μ e deviazione standard σ. Da essa si estraggano tutti i possibili campioni casuali della stessa dimensione n, e siano $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ la media e la deviazione standard della distribuzione della media campionaria. È vero che si ha:
$$\mu_{\overline{X}} = \mu\,,\quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\ ?$$
7. Si consideri un'indagine riguardante la presenza di un determinato carattere in una popolazione di dimensione N. Siano p la percentuale della popolazione che presenta quel carattere e σ la deviazione standard e siano inoltre $\mu_{\overline{p}}$ e $\sigma_{\overline{p}}$ la percentuale media e la deviazione standard della distribuzione della media campionaria, riferita a campioni di dimensione n. È vero che si ha:
$$\mu_{\overline{p}}=p\,,\quad \sigma_{\overline{p}}=\sqrt{\frac{p(1-p)}{n}}\ ?$$
8. Considerata la distribuzione della media campionaria di una popolazione e ammesso che siano $\mu_{\overline{X}}$ e $\sigma_{\overline{X}}$ (eventualmente $\mu_{\overline{p}}$ e $\sigma_{\overline{p}}$) le sue caratteristiche, è vero che si assumono come stime di queste i seguenti valori:
$$\mu_{\overline{X}}=\overline{x}\,,\quad \sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}} \quad \text{oppure} \quad \mu_{\overline{p}}=f\,,\quad \sigma_{\overline{p}}=\sqrt{\frac{f(1-f)}{n}}$$
dove $\overline{x}$ (od f) ed s sono le caratteristiche di un campione casuale estratto dalla popolazione?
9. Che significato bisogna attribuire all'espressione "Il livello di confidenza affinché la media μ (o, eventualmente, la percentuale p) di una data popolazione cada in un determinato intervallo di confidenza è 90%"?
10. Cos'è una stima puntuale? Cosa una stima intervallo?

11. Qual è l'enunciato del teorema centrale del limite?
12. Cos'è l'errore standard della media?
13. Si vuole indagare sulla presenza di un determinato carattere in una popolazione, ma non si hanno informazioni sulla percentuale di individui della popolazione che presenta quel carattere. Per questo è necessario che il fenomeno che si studia abbia la massima variabilità possibile. Come fare?

RISPOSTE.

1. È falso. Un ragionamento induttivo è quel modo di ragionare che, partendo da statistiche riferite ad un campione rappresentativo di una popolazione, permette di stimare i parametri corrispondenti riferiti alla popolazione intera.
2. Un parametro è una caratteristica del collettivo statistico ed è costante, mentre una statistica è una caratteristica del campione di riferimento e varia da campione a campione.
3. La risposta è negativa. Il campione è rappresentativo della popolazione intera da cui è tratto se è del tutto casuale, cioè se tutti gli individui che compongono la popolazione considerata hanno uguale probabilità di far parte del campione.
4. No. La dimensione di un campione è il numero degli individui che lo compongono.
5. Estratti da una popolazione statistica tutti i possibili k campioni di data dimensione e considerate le medie di questi campioni (medie campionarie), la distribuzione della media campionaria è la variabile statistica che ad ogni campione associa la sua media con la probabilità 1/k di essere estratto.
6. No. Solo la prima relazione è vera. Se però $n<N/20$, la seconda vale con buona approssimazione.
7. No. Solo la prima relazione è vera. Se però $n<N/20$, la seconda vale con buona approssimazione.
8. È vero.
9. Si intende affermare che, se si estraggono campioni casuali della popolazione, tutti della stessa dimensione, la percentuale di essi che presenta un intervallo di confidenza nel quale è contenuta la media μ (o la percentuale p) della popolazione è pari al 90%.
10. Una stima puntuale di un parametro della popolazione statistica su cui s'indaga è una stima espressa solamente da un valore numerico. Una stima intervallo si riferisce invece ad un intervallo di valori entro il quale può cadere il parametro su cui s'indaga ed alla probabilità che vi cada.
11. Questo è l'enunciato del **teorema centrale del limite**: Posto di estrarre campioni casuali della stessa dimensione n da una popolazione di media μ e deviazione standard σ, quando n è sufficientemente grande ($n\geq 30$) la distribuzione della media campionaria si approssima alla distribuzione normale con media μ e deviazione standard $\sigma/\sqrt{n}$.
12. L'errore standard della media non è altro che la deviazione standard campionaria, cioè è la quantità $\sigma_{\overline{X}}=\sigma/\sqrt{n}$, dove σ è la deviazione standard della popolazione ed n è la sua dimensione.
13. Se si indica con p la percentuale di individui che presenta il carattere su cui si indaga, bisogna supporre che il prodotto $p(1-p)$ sia il massimo possibile. Questo accade quando $p=0{,}5$.

LETTURA

Locale e globale.

I termini locale e globale non hanno bisogno di spiegazioni. Diversamente da altri binomi (finito/infinito, discreto/continuo, deterministico/stocastico, eccetera) in cui c'è una contrapposizione fra i due termini, in questo caso si può dire che c'è una vera e propria "alleanza" fra i termini medesimi. È infatti

possibile imbattersi in situazioni in cui dalla conoscenza "locale" di un "oggetto matematico" si può risalire alla conoscenza "globale" dello stesso e questo è un risultato notevole, essendo spesso difficile ottenere direttamente la conoscenza globale mentre può essere più semplice ottenere quella locale. Questo non accade sempre.

Un esempio, ancorché banale: la conoscenza di un punto di una retta è poco per conoscere la retta, ma già la conoscenza di due punti distinti di una retta determina completamente la retta.

Vediamo altri esempi in cui dalla conoscenza locale si risale a quella globale.

Viene in mente, anzitutto, il procedimento seguito da Eratostene per misurare la circonferenza della Terra. Abbiamo avuto modo di descriverlo in passato, per cui ci limitiamo ad un breve cenno. Eratostene non conosceva la lunghezza di tale circonferenza. Calcolò allora in modo ingegnoso la lunghezza e l'ampiezza di un arco della circonferenza e, con una semplice proporzione, da questa lunghezza risalì a quella della circonferenza.

Un'altra situazione riguarda i polinomi in una indeterminata a coefficienti reali. Abbiamo fatto vedere a suo tempo come la conoscenza locale del polinomio implichi, sotto opportune condizioni, quella globale.

Anche lo studio delle funzioni avviene spesso partendo da alcune informazioni locali per ottenere l'andamento globale del grafico della funzione.

Una situazione più raffinata si ha con un sistema assiomatico, come la geometria. La sua struttura "globale" è determinata da quella "locale" ossia dai pochi assiomi che ne sono alla base.

Se vogliamo, anche il metodo induttivo fa registrare una "collaborazione" fra il locale, vale a dire il campione su cui s'indaga, e il globale, cioè l'universo statistico da cui il campione è estratto. In questo caso si vede come ragionando sulle caratteristiche del campione si possano stimare quelle dell'universo che rappresenta.

Un esempio suggestivo di alleanza locale/globale è costituita da certe figure geometriche nelle quali lo stesso motivo si ripete su scale diverse a livello locale e globale: si tratta dei **frattali**. Un frattale è infatti un oggetto geometrico tale che se si ingrandisce un suo particolare si ottiene una figura che sembra essere l'oggetto di partenza.

Il termine "frattale" deriva dal latino *fractus* (spezzato, frazionato) [10]. Fu coniato nel 1975 da **Benoît Mandelbrot** (1924-2010), matematico polacco naturalizzato francese, che lo usò nel suo libro intitolato per l'appunto *Gli oggetti frattali*, in cui tratta abbondantemente dell'argomento, corredandolo di immagini suggestive (a titolo di esempio: Fig. 10, ma anche Fig. 11, benché quest'ultima, ad onor del vero, sia una creazione del matematico polacco **Waclaw Sierpinski**, 1882-1969, ed è conosciuta come *triangolo di Sierpinski*).

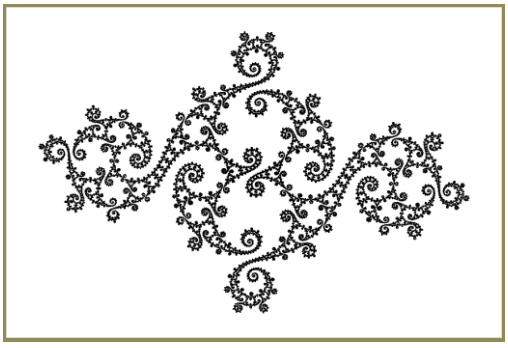

FIG. 10

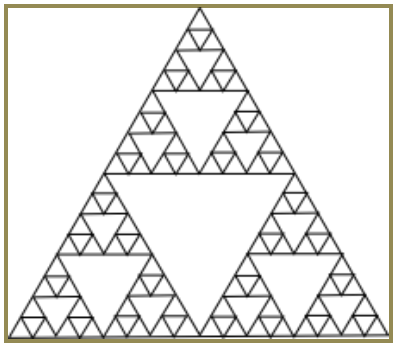

FIG 11

[10] Il nome ha a che fare con la dimensione di un frattale, che è appunto un numero frazionario.

Quantunque in natura esistano innumerevoli “oggetti frattali” (dai fiocchi di neve ai cavolfiori agli abeti), per far capire bene cosa intendiamo dire quando affermiamo che “lo stesso motivo si ripete a livello locale ed a livello globale” facciamo riferimento ad una singolare linea, che è stata costruita quando ai frattali nessuno ancora pensava. Si tratta della ***curva di Kock*** [11], ideata nel 1904 e conosciuta per lo più perché è caratterizzata da un’anomalia: non ammette tangente in alcun suo punto. La sua costruzione è accennata nella figura 12 e dà l’idea della ripetitività all’infinito dello stesso motivo.

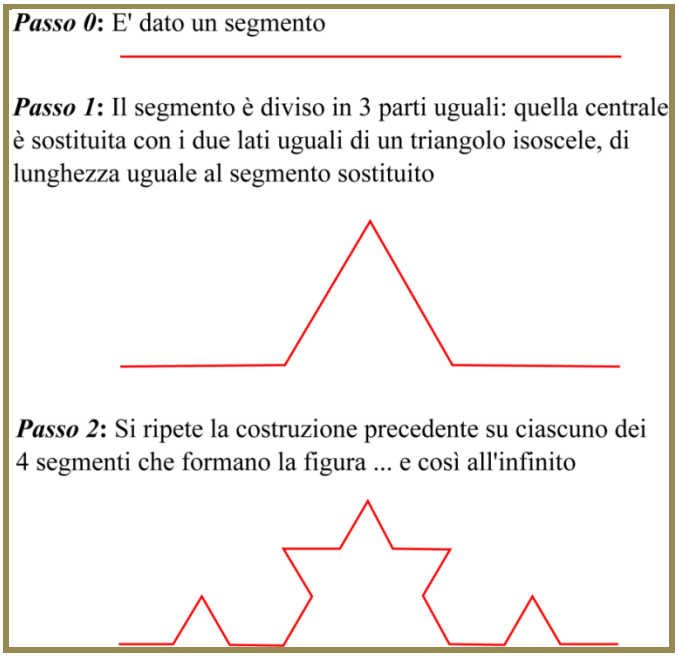

FIG. 12

Per concludere, pensiamo possa suscitare qualche interesse il seguente pensiero di Isaac Asimov (1920-1992), scrittore e divulgatore scientifico statunitense di origine russa [12]:

Io credo che il sapere scientifico abbia delle proprietà frattali, e che indipendentemente da quanto impariamo, quel che rimane, benché possa sembrare piccolo, è così infinitamente complesso come lo era il totale da cui siamo partiti.

[11] **Helge von Koch**, matematico svedese, 1870-1924.
[12] Fonte: Clifford Pickover, *La magia dei numeri*, RBA Italia, 2008, pag. 44.