DATI e PREVISIONI

Elenco delle attività

Conviene rispondere a caso?

Abilità	Conoscenze	Nuclei coinvolti	Collegamenti esterni
Spiegare il concetto e la funzione "variabile aleatoria". Utilizzare il teorema centrale del limite, approssimando le probabilità dalla distribuzione binomiale a quella normale. Determinare i parametri di una distribuzione ed usarli, eventualmente, anche per costruire la distribuzione gaussiana approssimante.	La distribuzione binomiale: il suo uso e le sue proprietà. La distribuzione normale: il suo uso e le sue proprietà.	Dati e previsioni Argomentare, congetturare, dimostrare.	Italiano Società civile

Contesto
Sociale: istruzione.

L'attività nasce dalla curiosità degli studenti circa la possibilità di affrontare con più o meno chances, avendo studiato in modo superficiale, un compito scritto a scuola formulato con domande a risposta multipla (ricordiamo che anche l'esame per la patente automobilistica e l'esame per il certificato di conoscenza della lingua inglese sono prove di valutazione di questi tipo).
In effetti la tentazione dello studente di "indovinare" la risposta esatta è molto forte, stimolato anche da considerazioni del mondo esterno. Nei mass media sono poi presenti molti giochi a premi in cui appare lampante il fatto che la fortuna gioca un ruolo predominante: questo può portare a pensare che rispondere bene è più facile di quanto si possa credere.

Descrizione dell'attività:
Si cerca di far capire, attraverso precise considerazioni probabilistiche, che cercare di rispondere a caso, sia in un compito in classe costituito da domande a risposta multipla, sia nel caso di test attitudinali nei momenti di assunzione al lavoro, non è sempre "pagante" dal punto di vista del risultato.
Supponiamo di dare una serie di 10 domande a risposta multipla, ciascuna con 4 possibili risposte, una sola delle quali è corretta. Si suppone anche che ogni domanda sia *indipendente* dalle altre. Ci si può allora chiedere: *Segnando a caso le risposte, qual è la probabilità di rispondere esattamente a 3 domande? Qual è la probabilità di rispondere esattamente a meno di 3 domande? E a più di 8 ? Qual è il numero atteso di risposte esatte ?*
Per rispondere alle domande precedenti è opportuno *ricorrere* al modello binomiale che fornisce le probabilità di ottenere k successi su N prove indipendenti:

$$P(X = k) = \binom{N}{k} \cdot p^k (1-p)^{N-k}, \qquad \text{con } 0 \le k \le N,$$

dove p è la probabilità di avere un successo in ciascuna *prova* e X è il numero (incognito) di successi, cioè di volte in cui si realizza l'evento a cui siamo interessati, su N prove.

L'insegnante, anche con l'uso del foglio elettronico, può guidare gli studenti a rilevare le caratteristiche del modello binomiale, la cui corrispondente distribuzione è indicata con $X \sim Bi(N, p)$, al variare di N e p. Fa notare che: 1) per p = 0,5 la distribuzione è simmetrica per ogni valore di N; 2) al crescere di N la distribuzione tende ad assumere una forma a campana; 3) il valor medio di X è uguale a N·p; 4) la varianza di X è uguale a N·p·(1-p).
Con riferimento alla serie delle 10 domande, se si sceglie a caso una delle quattro risposte, la probabilità di indovinare la risposta esatta è $\frac{1}{4}$. La probabilità di *rispondere* esattamente (a caso) a 3 domande su 10 è fornita da:

$$P(X=3)=\binom{10}{3}\cdot\left(\frac{1}{4}\right)^3\left(1-\frac{1}{4}\right)^7\cong 0{,}25$$

La probabilità di rispondere esattamente a meno di 3 domande è:

$$P(X<3)=\sum_{k=0}^{2}\binom{10}{k}\cdot\left(\frac{1}{4}\right)^k\left(1-\frac{1}{4}\right)^{10-k}\cong 0{,}53$$

La probabilità di rispondere a più di 8 domande è:

$$P(X>8)=\sum_{k=9}^{10}\binom{10}{k}\cdot\left(\frac{1}{4}\right)^k\left(1-\frac{1}{4}\right)^{10-k}\cong 0{,}000030$$

Il numero atteso delle risposte esatte è: $N\cdot p=\frac{10}{4}=2{,}5$
L'insegnante può far notare che, se il numero N delle domande diventa grande, la valutazione della probabilità, usando la distribuzione binomiale, diventa difficile dal punto di vista computazionale. In questo caso ci aiuta la teoria che dice che è possibile usare la distribuzione normale al posto della distribuzione binomiale, purché p sia non eccessivamente piccolo né troppo grande (ad es. compreso fra 0,2 e 0,8). Allora, per un fissato valore del numero dei successi, si ha:

$$P(X=k_1)=\binom{N}{k_1}\cdot p^{k_1}(1-p)^{N-k_1}\cong\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{k_1-\mu}{\sigma}\right)^2}$$

dove $\mu = N\cdot p$; $\sigma=\sqrt{N\cdot p\cdot(1-p)}$.

Se il test è costituito da 300 domande ciascuna con 4 possibili risposte, ci possiamo chiedere:
Qual è la probabilità di rispondere a caso correttamente a 70 domande?
Qual è la probabilità di rispondere correttamente ad almeno 70 domande?
Qual è il numero minimo di risposte esatte affinché la probabilità di superare tale valore sia 0,23?

Per calcolare la probabilità di rispondere correttamente a 70 domande, usando l'approssimazione alla normale (e quanto detto nella Nota precedente), sarà sufficiente sostituire nella distribuzione di probabilità a X il valore 70, a μ il valore $N\cdot p = \frac{300}{4} = 75$ e a σ il valore $\sqrt{Np(1-p)} = 7{,}5$.
Sostituendo dunque in:

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ i valori assegnati, otteniamo P(X=70) = *f*(70) = 0.042

Per calcolare P(X ≥ 70) si usano le Tavole della distribuzione normale standardizzata in corrispondenza al valore $t = \frac{k - 0.5 - \mu}{\sigma}$, sostituendo a k il valore 70, a μ il valore 75 e a σ il valore 7,5. Si ottiene:

$P(X \geq 70) = P\left(Z \geq \frac{69{,}5 - \mu}{\sigma}\right) = P(Z \geq -0.73) = 1 - P(Z \leq -0.73) = 0.7673$.

Per l'ultima domanda il procedimento è inverso: nota una probabilità si vuol ricercare il valore corrispondente di X, numero di risposte corrette.
Sulle Tavole della distribuzione normale si ricerca il valore t in corrispondenza a:

$$P(Z \geq t) = 0{,}23$$

e si ricava dalla relazione $t = \frac{k - 0{,}5 - \mu}{\sigma}$ il valore di k.
Nel nostro esempio t = 0,74 per cui k=0,74σ+μ + 0,5 = 81,05.
Quindi il numero minimo di domande a cui rispondere correttamente è 81.

A questo punto siamo anche pronti a rispondere a domande del tipo:
Come opera una industria che cerca un responsabile del personale del tipo "capace come ce ne sono solo 4 su 10000 al mondo"?
Occorre impostare l'equazione P(X ≥ n) = 0,0004 in cui n rappresenta l'incognita, ovvero il numero minimo di domande alle quali dovrà rispondere correttamente il candidato per essere uno dei "4 su 10000".
Supponiamo che l'esperienza del passato ci dica che al nostro test attitudinale i candidati sanno rispondere mediamente bene a solo 100 domande su 300, con uno scarto quadratico medio pari a 10. Dunque abbiamo ancora bisogno delle Tavole della distribuzione normale standardizzata.
L'equazione da risolvere è: P(Z ≥ t) = 0,0004 in corrispondenza al valore $t = \frac{n - 0{,}5 - 100}{10}$.
Si ricava, come nella domanda precedente, il valore t = 3,38 per cui n = 3,38·σ + μ + 0,5 = 134,3.
Il nostro candidato, per essere preso in considerazione, dovrà dunque essere capace di rispondere ad almeno 134 domande su 300.

Nota per il docente. Da considerazioni di analisi matematica (teorema del valor medio integrale) sappiamo che possiamo scrivere l'ultima approssimazione. Essa ci permetterà di approssimare il valore cercato "direttamente" come il valore che la funzione di densità prende nel punto "numero dei successi".

La funzione di densità di probabilità del modello normale risulta:

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ con x reale,

al variare di μ e σ, in particolare partendo da μ=0 e σ=1.
Se si vuole calcolare per una binomiale la probabilità di un numero di successi compreso fra k_1 e k_2 usando l'approssimazione con la distribuzione normale si ha:

$$P(k_1 \le k \le k_2) = \sum_{k=k_1}^{k_2} \binom{N}{k} \cdot p^k (1-p)^{N-k} \cong \int_{k_1-\frac{1}{2}}^{k_2+\frac{1}{2}} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{k-\mu}{\sigma}\right)^2} dk$$

Se nell'integrale si effettua la sostituzione $t = \frac{k-\mu}{\sigma}$, l'integrale diventa:

$$\int_{\frac{k_1-0,5-\mu}{\sigma}}^{\frac{k_2+0,5-\mu}{\sigma}} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}t^2} \sigma\, dt$$

Questa è la valutazione numerica della probabilità, che la variabile normale *standardizzata* (cioè quella che ha media 0 e deviazione standard 1) assuma un valore nell'intervallo $\left(\frac{k_1-0,5-\mu}{\sigma}, \frac{k_2+0,5-\mu}{\sigma}\right)$.

Essa viene fornita da Tavole numeriche in quanto è noto, sempre dall'analisi matematica, che la funzione e^{-t^2} è una delle funzioni che non ammette primitiva esprimibile in termini elementari (questo è il motivo per cui su ogni libro di fisica, di statistica, di probabilità, ... sono riportate le Tavole della distribuzione normale).

Esempi di prove di verifica

1. Se si lancia 15 volte una moneta perfetta, qual è la probabilità di avere esattamente 11 Teste ?

2. Se si lancia 15 volte una moneta perfetta, qual è la probabilità di avere almeno 11 Teste ?

3. Quanti lanci di un dado si prevede di dover fare perché sia 0,95 la probabilità di avere almeno una volta il numero 6 ?

4.

a) In un test vi sono 150 domande. Per esperienza passata si sa che i candidati rispondono correttamente a 9 domande su 10. Mediamente quante risposte corrette darà un candidato?
b) In un test vi sono 150 domande. Per esperienza passata si sa che i candidati rispondono correttamente 2 volte su 3. In media, quante risposte corrette darà un candidato? E con quale scarto quadratico medio? In un test vi sono 150 domande con 5 risposte per ogni domanda. Se un candidato risponde a caso, a quante domande risponderà correttamente in media?

5. Un insegnante porta la sua classe in settimana bianca. Si sa che ogni studente ha la probabilità $\frac{1}{100}$ di farsi male sciando e che l'insegnante riesce a riportare incolumi tutti gli studenti a scuola 3 volte su 4. Si può dare una stima del numero di allievi che l'insegnante ha in quella classe ?

Campioni e "forchette"

Abilità	Conoscenze	Nuclei coinvolti	Collegamenti esterni
Spiegare il concetto e la funzione "variabile aleatoria". Costruire un campione casuale semplice, data una popolazione. Stimare una proporzione, o la media di una popolazione, attraverso una consapevole costruzione di stime puntuali e intervalli di confidenza.	Campionamento casuale semplice e distribuzione campionaria della proporzione e della media. Stima puntuale e intervallo di confidenza per la proporzione e per la media.	Dati e previsioni Argomentare, congetturare, dimostrare Laboratorio di matematica	Italiano Società civile Storia

Contesto
Sociale

I mezzi di comunicazione di massa ci hanno abituati all'uso dei sondaggi i cui risultati dipendono da un campione. Il ricorso ad esempi tratti dalla vita quotidiana motiverà gli studenti al tema di questa attività che intende sviluppare semplici metodi di campionamento al fine di evidenziarne le caratteristiche principali e di comprenderne i limiti con particolare riferimento alla stima della proporzione.

Descrizione dell'attività
Un campione è un sottoinsieme dell'intera popolazione. Le informazioni estratte vengono utilizzate per trarre delle conclusioni più generali riferite a tutta la popolazione di cui il campione osservato è solo una parte.
La validità delle conclusioni tratte e il controllo sull'errore che si commette usando una parte anziché il tutto dipendono dal modo in cui l'estrazione è effettuata.
I motivi che giustificano l'uso di un campione piuttosto che l'intera popolazione possono essere diversi:

a) la popolazione può essere molto vasta: risulta troppo lungo analizzare tutte le unità statistiche (un sondaggio di opinioni viene fatto di solito su un insieme ristretto di individui poiché se ne vogliono conoscere gli esiti con rapidità);
b) le misure possono essere distruttive; può essere il caso di un controllo delle misure di affidabilità di un componente elettronico;
c) il costo della rilevazione totale della popolazione può essere esorbitante.

In tutti questi casi è necessario effettuare un campionamento, cioè l'estrazione di un campione. Se l'estrazione è casuale si parla di campione casuale. L'estrazione avviene in modo che ogni elemento della popolazione abbia la stessa probabilità di tutti gli altri di essere estratto. Ovviamente campione casuale non significa campione preso "a caso", ma ottenuto rispettando delle leggi probabilistiche nella scelta. Ad esempio: campione formato da elementi scelti in modo equiprobabile dalla popolazione di partenza. In questo caso il calcolo delle probabilità è di aiuto nella costruzione della variabile campionaria e consente di quantificare l'errore che si commette estendendo le informazioni del campione all'intera popolazione.

Nel campionamento è facile commettere l'errore di ritenere casuale un campione che invece non lo è. Ad esempio, se si vuole effettuare un sondaggio di opinioni sulle abitudini sportive degli abitanti di una città e si decide di scegliere un campione di 200 abitanti, risulta errato scegliere le persone:

- da un elenco telefonico, in quanto verrebbero esclusi tutti coloro che non compaiono nell'elenco;
- tra coloro che transitano su una strada, in quanto verrebbero esclusi i frettolosi, mentre si tenderebbe a raccogliere dati da coloro che sembrano più educati o ben disposti a chiacchierare.

Per ottenere un campionamento realmente casuale si può invece procedere nel seguente modo:

- si assegna ad ogni abitante della città un numero progressivo;
- si pongono in un'urna tutti i numeri assegnati;
- si estraggono tanti numeri (in questo caso 200) quante sono le persone da intervistare;
- si intervistano le persone che corrispondono ai numeri estratti.

Procedendo in questo modo si può effettivamente parlare di campionamento casuale, in quanto tutti i numeri sono equiprobabili, e quindi tutte le persone hanno la stessa probabilità di essere intervistate. Gli schemi fondamentali di campionamento casuale sono due, in quanto il campionamento può essere effettuato con o senza ripetizione, ovvero con o senza reintroduzione (nella popolazione) dell'unità statistica estratta. Nel primo caso la stessa unità statistica può essere estratta più volte, mentre nel secondo caso ciascun elemento della popolazione può presentarsi una sola volta.

Nel caso di campionamento senza reintroduzione la composizione della popolazione varia da estrazione ad estrazione; tuttavia, quando la popolazione è infinita (molto vasta) e il campione è piccolo, il fatto che vi sia o meno reintroduzione perde importanza, in quanto la composizione della popolazione rimane pressoché costante nel corso delle estrazioni.

Un famoso episodio di campione distorto fa riferimento all'indagine condotta nel 1936 dalla rivista americana "Literary Digest" con l'obiettivo di prevedere l'esito delle elezioni presidenziali di quell'anno (si veda ad es. Brusati, 2003).

Gli autori del sondaggio inviarono per posta dei facsimili della scheda elettorale a oltre 10 milioni di persone, individuate attraverso gli elenchi telefonici e i registri automobilistici dell'epoca. Gli intervistati rimandarono indietro circa due milioni di risposte che, una volta elaborate, diedero come risultato il successo di A. London con circa il 60% delle preferenze. Le elezioni smentirono in modo clamoroso tali previsioni in quanto F.D. Rooswelt ottenne oltre il 61% delle preferenze degli elettori americani. L'esito reale dei risultati fu invece correttamente previsto dall'istituto di sondaggi "Gallup" che operò con un campione notevolmente inferiore.

Perché un campione di dimensioni notevolmente ridotte rispetto a quello della Literary Digest aveva centrato correttamente le previsioni mentre quello di dimensioni molto più ampie aveva fallito? Cosa era successo?

Il motivo di un tale insuccesso può essere ricondotto a due tipi di errore. Prima di tutto ci fu un *errore di copertura* (non rappresentatività) in quanto, avendo preso come riferimento gli elenchi telefonici e i registri automobilistici, si utilizzarono liste incomplete della popolazione e vennero trascurate le fasce più deboli e meno abbienti che votarono evidentemente in modo opposto a quella di chi disponeva di un telefono o era proprietario di una automobile. Il procedimento seguito, cioè, lasciò fuori dal campione una parte della popolazione con caratteristiche proprie e particolari: la popolazione meno abbiente. Un'altra distorsione del campione derivò da un *errore dovuto alle mancate risposte*: infatti dei 10 milioni di questionari inviati tornarono indietro solo due milioni (il 20%). Il campione delle risposte fu il risultato di una *autoselezione* notevole: cioè le manifestazioni di voto degli elettori che risposero al questionario non erano simili a quelle di coloro che non risposero.

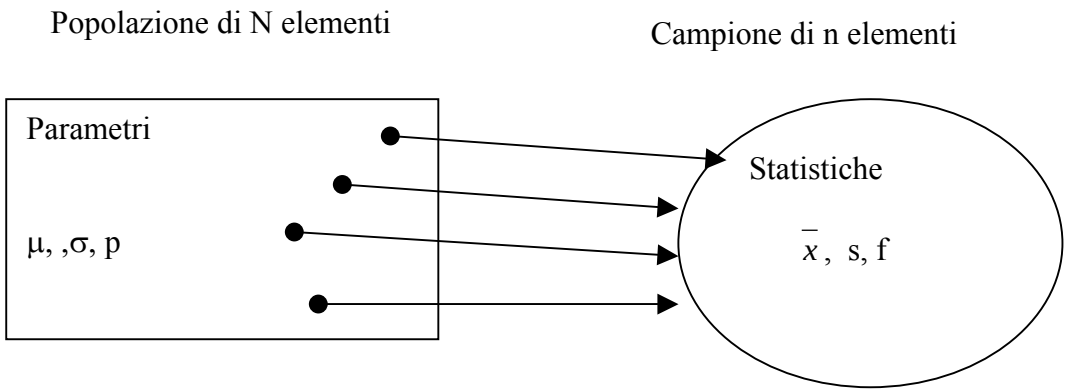
Il campione costruito da Gallup, pur di numerosità nettamente inferiore ma *scelto casualmente* (seguendo criteri probabilistici) tra l'intera popolazione, risultò più simile a tutta la popolazione rispetto alla formulazione di voto.

Può essere utile porre attenzione ai simboli e ai termini usati quando si parla di campionamento.
In particolare può essere utile ricordare che:

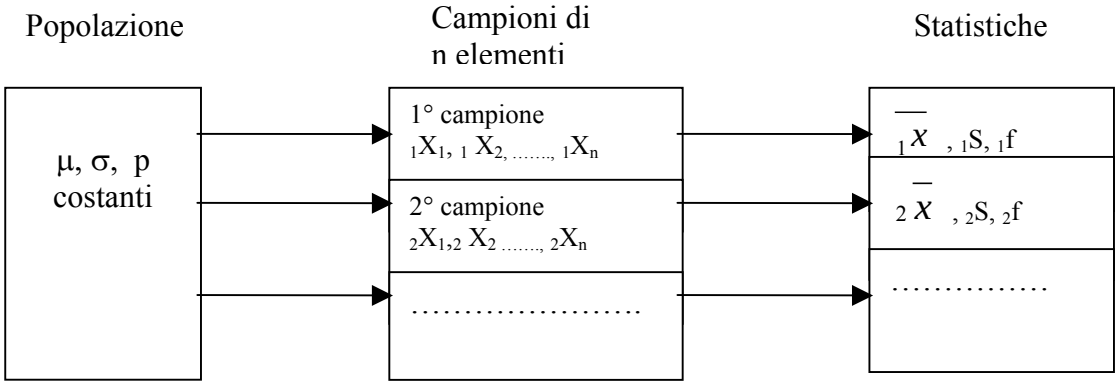
- la misura di una caratteristica "della" popolazione sarà chiamata *parametro*;
- la misura di una caratteristica "di un" campione sarà chiamata *statistica*.

	Popolazione	Campione
Definizione	Insieme di tutte le unità statistiche	Sottoinsieme di unità selezionate in modo casuale dalla popolazione
Caratteristica	Parametro *	Statistica **
media	μ	$\bar{x}$
deviazione standard	σ	s
varianza	σ^2	s^2
frequenza relativa	p	f
ampiezza	N	n
caratteristica generica	θ	t

* è una caratteristica della popolazione ed è costante
** è una caratteristica del campione e può variare da campione a campione



Fissata la popolazione, quanti campioni si possono estrarre da essa? Se si indica con N il numero di unità statistiche che formano la popolazione e con n quelle del campione, quando le estrazioni avvengono con il metodo della reintroduzione, i possibili campioni sono *le disposizioni con ripetizione* di N elementi di classe *n* e sono N^n.
Si considerino i dati di un campione: su di esso è possibile calcolare delle statistiche che possono variare al variare del campione. Infatti, mentre i parametri di una popolazione sono *valori numerici costanti*, le statistiche campionarie sono *variabili aleatorie (casuali.)*



I risultati della indagine su un campione si possono utilizzare o per descrivere le caratteristiche del campione osservato (analisi descrittiva della situazione) oppure per trarre informazioni su alcuni

aspetti della popolazione. Questo secondo utilizzo delle informazioni, che dal particolare passa al generale, prende il nome di inferenza statistica che è inferenza induttiva. In questa unità si analizzerà, sulla base di dati reali, la procedura di estrazione casuale semplice e alcune conseguenze ad essa connesse. Il database sul quale sono stati costruiti vari campioni è riportato nella tabella seguente e fa riferimento al sesso, al comune di residenza e al voto scritto di matematica ottenuto dagli studenti delle classi terze informatica dell'Itis "C. Zuccante" di Mestre-Ve.

ID	Sesso	Comune	Voto Mat.	ID	Sesso	Comune	Voto Mat.	ID	Sesso	Comune	Voto Mat.	ID	Sesso	Comune	Voto Mat.
1	M	Altro	5	44	M	Altro	5	87	M	Altro	4	130	M	Altro	7
2	M	Venezia	3	45	M	Altro	5	88	M	Altro	5	131	M	Venezia	6
3	F	Altro	6	46	F	Altro	4	89	M	Altro	8	132	M	Altro	7
4	M	Altro	5	47	M	Venezia	5	90	M	Altro	5	133	M	Altro	7
5	M	Altro	3	48	M	Venezia	5	91	M	Venezia	6	134	M	Venezia	7
6	F	Altro	7	49	F	Venezia	6	92	M	Venezia	5	135	M	Venezia	5
7	F	Altro	5	50	F	Venezia	6	93	M	Venezia	8	136	M	Venezia	7
8	F	Altro	6	51	M	Venezia	5	94	M	Venezia	6	137	F	Venezia	5
9	M	Venezia	6	52	M	Altro	4	95	M	Venezia	3	138	M	Venezia	5
10	M	Venezia	5	53	M	Altro	4	96	M	Venezia	5	139	M	Venezia	6
11	F	Venezia	6	54	M	Altro	4	97	M	Venezia	5	140	M	Venezia	6
12	M	Venezia	7	55	M	Altro	7	98	F	Venezia	6	141	M	Venezia	6
13	M	Altro	5	56	M	Venezia	4	99	M	Venezia	6	142	M	Venezia	5
14	M	Altro	5	57	M	Altro	5	100	M	Venezia	3	143	F	Venezia	7
15	M	Venezia	6	58	M	Venezia	5	101	F	Venezia	4	144	M	Altro	6
16	M	Venezia	9	59	M	Altro	5	102	M	Altro	4	145	M	Venezia	6
17	M	Altro	5	60	M	Altro	4	103	M	Venezia	4	146	M	Venezia	7
18	M	Altro	4	61	M	Altro	6	104	M	Venezia	6	147	M	Altro	9
19	M	Altro	5	62	M	Altro	6	105	M	Venezia	4	148	M	Altro	7
20	M	Venezia	7	63	M	Altro	4	106	M	Venezia	4	149	M	Altro	5
21	M	Altro	6	64	M	Altro	4	107	M	Venezia	8	150	M	Altro	6
22	M	Venezia	7	65	M	Venezia	2	108	M	Venezia	5	151	M	Venezia	7
23	M	Venezia	8	66	M	Venezia	3	109	M	Venezia	8	152	M	Venezia	5
24	M	Altro	7	67	M	Altro	7	110	M	Altro	7	153	M	Altro	5
25	M	Altro	5	68	M	Altro	7	111	F	Venezia	7	154	M	Venezia	6
26	M	Altro	9	69	M	Altro	5	112	F	Venezia	3	155	M	Altro	5
27	M	Altro	5	70	M	Venezia	5	113	M	Altro	7	156	M	Altro	6
28	M	Altro	4	71	M	Altro	4	114	M	Altro	7	157	M	Venezia	7
29	M	Venezia	6	72	M	Altro	5	115	M	Altro	9	158	M	Altro	6
30	M	Altro	6	73	M	Altro	3	116	M	Venezia	5	159	M	Venezia	6
31	M	Venezia	5	74	M	Venezia	6	117	M	Venezia	5	160	M	Altro	6
32	M	Altro	9	75	M	Venezia	4	118	M	Venezia	5	161	M	Altro	6
33	M	Altro	6	76	M	Altro	4	119	M	Venezia	5	162	M	Altro	7
34	M	Altro	4	77	M	Venezia	6	120	M	Venezia	5	163	M	Venezia	5
35	F	Altro	6	78	M	Altro	3	121	F	Venezia	5	164	M	Venezia	7
36	M	Altro	5	79	M	Altro	6	122	F	Venezia	8	165	M	Altro	5
37	F	Venezia	4	80	M	Altro	2	123	M	Altro	4	166	M	Altro	7
38	M	Venezia	6	81	M	Venezia	4	124	M	Altro	8	167	M	Altro	4
39	M	Altro	6	82	M	Venezia	3	125	M	Altro	5	168	M	Altro	8
40	F	Altro	4	83	M	Venezia	4	126	F	Altro	7	169	M	Altro	8
41	M	Altro	4	84	M	Altro	5	127	F	Altro	6	170	M	Altro	5
42	M	Altro	6	85	M	Venezia	5	128	M	Venezia	6	171	M	Altro	5
43	M	Altro	6	86	M	Venezia	7	129	F	Venezia	5	172	M	Altro	5

Tabella 1

Di questa popolazione di numerosità N=172 sono state rilevate le seguenti caratteristiche: Sesso, Comune di residenza e Voto scritto in matematica al primo quadrimestre, di cui vengono forniti alcuni indici sintetici:

Sesso		
		p_i
Maschi	150	0,87
Femmine	22	0,13

Comune di residenza		
		p_i
Venezia	81	0,47
Altro	91	0,53

Voto scritto in matematica

Voto matematica	media	Dev. Stand.	Varianza	Min	q1	mediana	q3	max
1° quadrimestre	5,53	1,41	1,99	2	5	5	6	9

Il grafico di Figura 1 rappresenta la distribuzione dei voti di matematica.

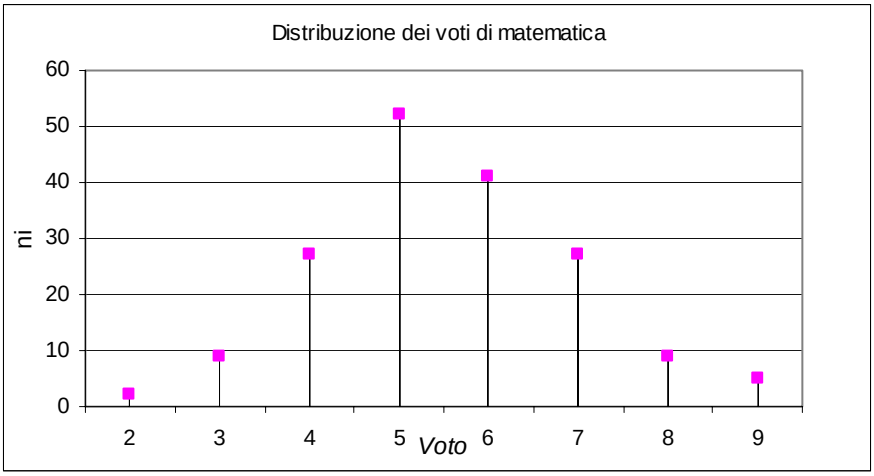


Figura 1

A questo punto l'insegnante invita gli studenti ad estrarre campioni casuali semplici usando il calcolatore. Di questi verranno calcolati gli indici sintetici già studiati per la popolazione.
La Tabella 2 indica il procedimento che consente di ricavare dal database di Tabella 1 un campione casuale di 10 elementi, per rilevare ad esempio il voto scritto in matematica (VotoMat). Il database deve essere disposto verticalmente e la prima colonna deve contenere il campo Id, la seconda colonna il campo Sesso, la terza il campo Comune e la quarta il campo Voto Mat. Nella colonna A è stata impostata la funzione del foglio elettronico che consente di generare un numero casuale compreso tra 1 e 172 e che individua, nella prima colonna del database degli studenti, il numero identificativo dello studente che entra a far parte del campione. Nella colonna B è stata impostata la funzione di ricerca "CERCA.VERT()", che ricerca nella prima colonna del database, indicato da "Foglio1!$A$1:$G$173", il valore numerico estratto dalla funzione in colonna A, individuando così una riga del database di Tabella 1; tale funzione restituisce il valore della colonna 4 nella corrispondente riga, ossia il voto scritto di matematica dell'alunno estratto.

	A	B
1	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A1;Foglio1!$A$1:$G$173;4)
2	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A2;Foglio1!$A$1:$G$173;4)
3	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A3;Foglio1!$A$1:$G$173;4)
4	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A4;Foglio1!$A$1:$G$173;4)
5	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A5;Foglio1!$A$1:$G$173;4)
6	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A6;Foglio1!$A$1:$G$173;4)
7	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A7;Foglio1!$A$1:$G$173;4)
8	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A8;Foglio1!$A$1:$G$173;4)
9	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A9;Foglio1!$A$1:$G$173;4)
10	=INT(CASUALE()*172)+1	=CERCA.VERT(Foglio3!A10;Foglio1!$A$1:$G$173;4)

Tabella 2

Con lo stesso procedimento sostituendo i numeri 2 e 3 nell'ultimo parametro della funzione CERCA.VERT(), è possibile ricavare rispettivamente il Sesso e il Comune dell'alunno selezionato.

Si ottiene la Tabella 3.

Campione di 10 elementi estratto casualmente dal database di Tabella 1

Id	Sesso	Comune	voto
14	M	Altro	5
126	F	Altro	7
131	M	Venezia	6
3	F	Altro	6
46	F	Altro	4
40	F	Altro	4
57	M	Altro	5
97	M	Venezia	5
28	M	Altro	4
119	M	Venezia	5

Tabella 3

Gli indici calcolati su questo campione sono i seguenti:

Sesso

		f_i
Maschi	6	0,60
Femmine	4	0,40

Comune di residenza

		f_i
Venezia	3	0,30
Altro	7	0,70

voto matematica 1° quadrimestre	media	Dev. Stand.	Varianza	Min	q1	mediana	q3	max
	5,10	0,99	0,98	4	4,25	5	5,75	7

L'insegnante chiede: la proporzione dei maschi nel campione è uguale a quella della popolazione? E la proporzione dei residenti nel comune di Venezia? La media dei voti nel campione è uguale a quella della popolazione?
L'insegnante chiede infine: in una successiva ripetizione dell'esperimento si otterranno gli stessi risultati o risultati diversi?

Se la risposta a tutte le domande è "no", perché si verifica ciò? Si è forse errato nel processo di estrazione? E' sempre possibile ripetere il percorso già effettuato. L'insegnante mostra il procedimento per produrre 10 campioni di numerosità 10.
Inizialmente si procede come indicato per la scelta di un unico campione. Poi mediante la funzione *Copia/Incolla* sono stati formati 10 campioni di numerosità 10 e i risultati dei caratteri osservati sono riportati in una tabella di sintesi. Per riprodurre tutti i campioni estratti conviene copiare i dati generati nel foglio precedente e incollarli nella nuova tabella attraverso l'opzione *Copia/Incolla speciale/Solo valori* in quanto il foglio elettronico modifica i numeri calcolati ogni volta che, in modo diretto o indiretto, viene investita la funzione Casuale().
Nella Tabella 4 sono riportati per ogni campione il valore della proporzione dei Maschi, della proporzione dei residenti nel comune di Venezia e la media del voto in matematica. Nelle ultime due colonne sono riportati media e varianza degli indici trovati.

Campione	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**media**	**Var.**
Prop. M	0,6	0,8	0,9	0,9	1	0,8	0,9	0,7	0,6	0,8	0,80	0,02
Prop. Ve	0,3	0,4	0,4	0,5	0,6	0,8	0,7	0,2	0,6	0,6	0,51	0,03
Media voto	5,1	6	5,9	5,7	4,7	5,7	5,9	5,2	5,3	5,4	5,49	0,18

Tabella 4

L'insegnante fa osservare che i valori trovati nei diversi campioni variano. In particolare per la proporzione dei residenti nel comune di Venezia fa riportare i dati come nel grafico di Figura 2 inserendo anche il valore della proporzione per l'intera popolazione. Fa inoltre notare che il valore medio delle proporzioni campionarie, che nei dieci campioni estratti vale 0,51, non si discosta di molto dal valore della proporzione nella popolazione che è 0,47.

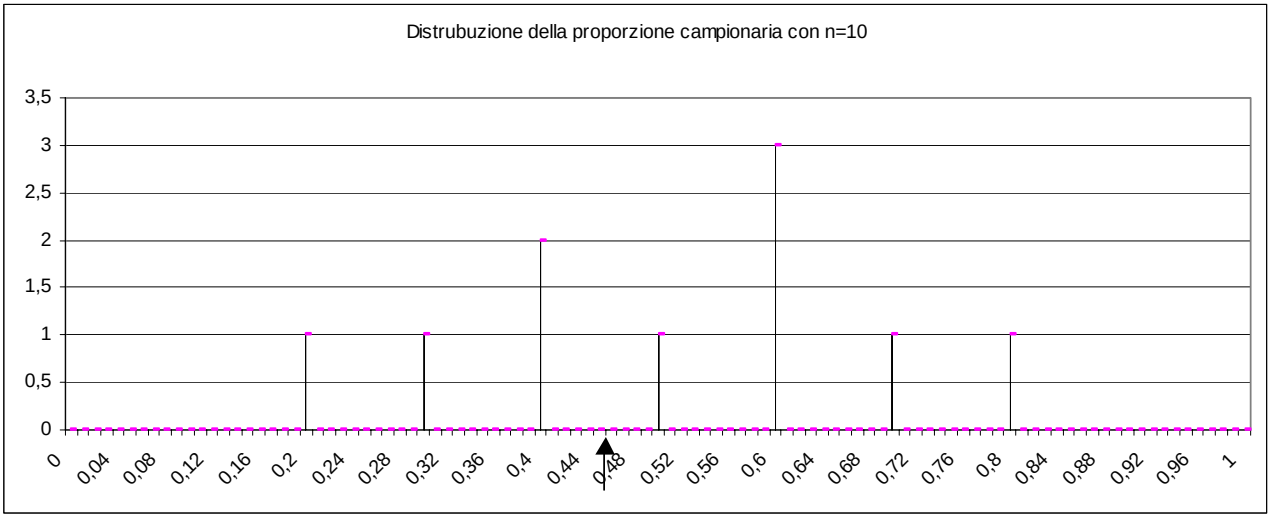


Figura 2

Analoghi ragionamenti potrebbero essere fatti per gli altri caratteri.

Cosa succede se la numerosità campionaria aumenta? Si porti ad esempio la numerosità a 40. Allora il procedimento già seguito dà luogo alla Tabella 5.

Campione	1	2	3	4	5	6	7	8	9	10	media	varianza
Prop. M	0,73	0,95	0,83	0,88	0,85	0,88	0,90	0,93	0,90	0,83	0,87	0,004
Prop. Ve	0,63	0,45	0,53	0,48	0,45	0,43	0,55	0,35	0,45	0,53	0,48	0,006
Media voto	5,55	5,18	5,13	5,58	5,48	5,35	5,58	5,68	5,53	5,48	5,45	0,033

Tabella 5

Nel grafico di Figura 3 è riportata la distribuzione della proporzione campionaria dei residenti nel comune di Venezia per i 10 campioni ottenuti con l'indicazione della proporzione per l'intera popolazione.

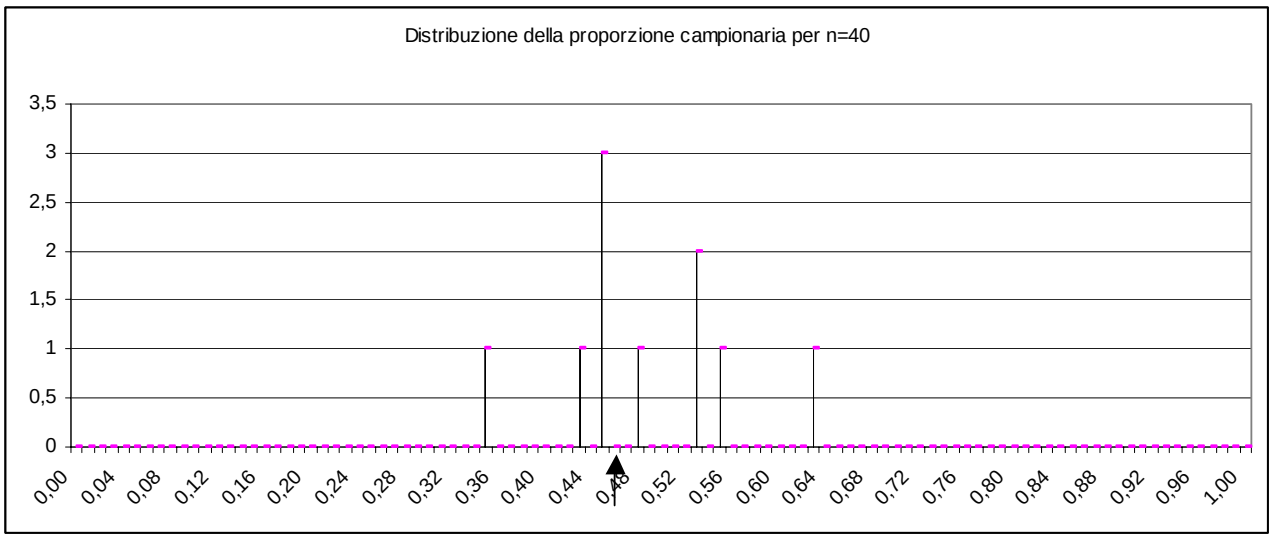


Figura 3

L'insegnante invita gli studenti a confrontare i due grafici e le tabelle corrispondenti e pone alcune domande: in quale caso si presenta maggior variabilità nei risultati ottenuti? In quale delle due situazioni la media delle proporzioni è più vicina alla proporzione della popolazione? Se si aumentasse il numero dei campioni estratti di numerosità 10, quale valore medio delle proporzioni campionarie ci si aspetterebbe di trovare?
Analoghi ragionamenti possono essere fatti per gli altri caratteri di tipo sia qualitativo sia quantitativo.
L'insegnante chiede se il risultato della proporzione campionaria può essere preso come valore della proporzione nella popolazione. Se la risposta è sì, con quale errore? E' diversa la risposta al variare di n?
L'insegnante ricorda ancora agli studenti il significato della deviazione standard. Chiede come interpreterebbero, con riferimento ai 10 campioni di numerosità 10 e alla proporzione dei residenti nel comune di Venezia, l'intervallo [0,51 –0,17; 0,51+ 0,17] e l'intervallo [0,60 –0,17; 0,60+ 0,17], con riferimento al decimo campione.
E' forse questo il significato della "forchetta" di cui parlano spesso i mass media?

Piccoli campioni crescono

Abilità	Conoscenze	Nuclei coinvolti	Collegamenti esterni
Spiegare il concetto e la funzione "variabile aleatoria". Costruire un campione casuale semplice, data una popolazione. Stimare una proporzione, o la media di una popolazione, attraverso una consapevole costruzione di stime puntuali e intervalli di confidenza.	Campionamento casuale semplice e distribuzione campionaria della proporzione e della media. Stima puntuale e intervallo di confidenza per la proporzione e per la media.	Dati e previsioni Argomentare, congetturare, dimostrare Laboratorio di matematica	Italiano Società civile

Contesto
Sociale: istruzione

I mezzi di comunicazione di massa ci hanno abituati all'uso dei sondaggi i cui risultati dipendono da un campione. Il ricorso ad esempi tratti dalla vita quotidiana motiverà gli studenti al tema di questa attività che intende sviluppare semplici metodi di campionamento al fine di evidenziarne le caratteristiche principali e di comprenderne i limiti con particolare riferimento alla stima della media campionaria.

Descrizione dell'attività
L'insegnante ricorda agli studenti quanto appreso in precedenza riguardo ai concetti di statistica descrittiva e al loro uso nell'analisi di un fenomeno: distribuzioni di frequenza, rappresentazioni grafiche, indici di sintesi quali i valori medi e le misure di variabilità. Ricorda inoltre che l'osservazione di una caratteristica può essere fatta sull'intera popolazione o su una sua parte. Come si può scegliere un campione di elementi dalla popolazione affinché i risultati da esso ottenuti siano trasferibili all'intera popolazione? E' possibile quantificare l'errore che si commette? Quando il campione è casuale, ossia scelto secondo i principi del calcolo delle probabilità, esiste la possibilità di quantificare e in qualche modo limitare l'errore e di mostrare come esso dipenda dalla numerosità campionaria (n). Si ribadisce l'importanza della scelta casuale del campione e come sia possibile fidarsi del campione per fornire informazioni sulla popolazione solo tenendo conto della probabilità.
Con i dati relativi alle valutazioni ottenute da N=6 studenti in un compito fatto in classe, si può avviare una attività sul campionamento, sulle caratteristiche dei dati che lo compongono e sulle statistiche campionarie da esso derivanti.

Studente	voti
A	8
B	6
C	4
D	5
E	6
F	7
media	6
varianza	1,667

A partire dai dati osservati l'insegnante invita gli studenti a creare tutti i possibili campioni casuali di n=2 elementi con la tecnica del campionamento semplice (con reinserimento). Quanti sono? Chiaramente sono N^n e quindi nel nostro caso sono $6^2=36$.
Sui campioni estratti si calcola la media dei voti. L'insegnante fa osservare come la media cambi al variare del campione. Una osservazione analoga era già stata fatta in precedenza? (si invita a vedere l'attività: Campioni e "forchette"). Si può costruire la distribuzione delle medie campionarie? L'insegnante invita gli studenti a riflettere su quale potrebbe essere la distribuzione di tale variabile. Quale sarà la media della distribuzione delle medie campionarie? E la varianza? L'insegnante conduce gli studenti a predisporre un prospetto analogo alla Tabella 1 che contiene lo spazio campionario relativo all'esperimento descritto sopra. In ogni casella vi è un possibile campione di numerosità 2 con i voti degli studenti estratti.

Campioni	A	B	C	D	E	F
A	(8; 8)	(8; 6)	(8; 4)	(8; 5)	(8; 6)	(8; 7)
B	(6; 8)	(6; 6)	(6; 4)	(6; 5)	(6; 6)	(6; 7)
C	(4; 8)	(4; 6)	(4; 4)	(4; 5)	(4; 6)	(4; 7)
D	(5; 8)	(5; 6)	(5; 4)	(5; 5)	(5; 6)	(5; 7)
E	(6; 8)	(6; 6)	(6; 4)	(6; 5)	(6; 6)	(6; 7)
F	(7; 8)	(7; 6)	(7; 4)	(7; 5)	(7; 6)	(7; 7)

Tabella 1

Nella tabella seguente sono riportate le medie campionarie conseguenti ai dati della Tabella 1.

Medie	A	B	C	D	E	F
A	8	7	6	6,5	7	7,5
B	7	6	5	5,5	6	6,5
C	6	5	4	4,5	5	5,5
D	6,5	5,5	4,5	5	5,5	6
E	7	6	5	5,5	6	65
F	7,5	6,5	5,5	6	6,5	7

Tabella 2

L'insegnante invita a costruire la distribuzione della media campionaria, a calcolare gli indici sintetici e a farne la rappresentazione grafica

La Tabella 3 indica il procedimento da seguire.

Distribuzione della media campionaria dei voti per campioni di 2 elementi da una popolazione di 6

$\overline{x}_i$	n_i	$\overline{x}_i \cdot n_i$	$\overline{x}_i^2 \cdot n_i$
4	1	4	16
4,5	2	9	40,5
5	5	25	125
5,5	6	33	181,5
6	8	48	288
6,5	6	39	253,5
7	5	35	245
7,5	2	15	112,5
8	1	8	64
Tot	36	216	1326
$M(\overline{X})$	6		
$Var(\overline{X})$	0,833		

Tabella 3

Il grafico di Figura 1 riporta la rappresentazione grafica della distribuzione della media campionaria

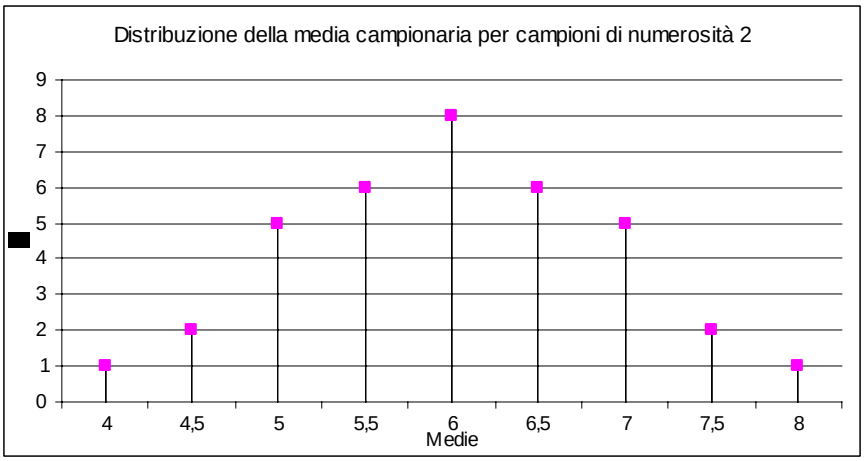


Figura 1

Quali informazioni è possibile ricavare dal grafico?

L'insegnante fa notare che:

a) la media della distribuzione della media campionaria è uguale alla media dei voti nella popolazione;
b) la varianza è minore;
c) la distribuzione della media campionaria è simmetrica rispetto al suo valore medio.

Perché le due medie sono uguali? Ciò si deve verificare sempre? Perché la varianza della media campionaria è minore di quella della popolazione? Ciò si deve verificare sempre? La forma del grafico assomiglia a qualche distribuzione nota?

L'insegnante fa notare che la distribuzione della media campionaria $\overline{X}$ possiede alcune proprietà valide qualsiasi sia il modello che descrive la caratteristica X nella popolazione di riferimento.

Tali proprietà affermano che: $M(\overline{X})=\mu$ e $Var(\overline{X})=\frac{\sigma^2}{n}$.

I simboli introdotti rappresentano rispettivamente:

M	Operatore valor medio
Var	Operatore varianza
$\overline{X}$	Variabile Media campionaria
μ	Media della popolazione
σ^2	Varianza della popolazione

L'insegnante invita gli studenti a verificare che le proprietà sono vere anche nello spazio campionario trattato.
L'insegnante propone di studiare sperimentalmente, col metodo della simulazione, la distribuzione della media campionaria su un certo numero di campioni di numerosità 2 estratti dagli studenti dai dati di origine relativi alle valutazioni ottenute da N = 6 studenti in un compito fatto in classe.
Ogni studente preleva un campione, con reinserimento di 2 elementi per 4 volte, e scrive le caratteristiche osservate: vi sono elementi ripetuti fra i 2 estratti? Osservando più sequenze cosa le diversifica?
Tale estrazione può essere fatta sperimentalmente inserendo sei dischi numerati da 1 a 6 (quelli della tombola) in un'urna, oppure ricorrendo al foglio elettronico e alla funzione Casuale() che restituisce un numero casuale distribuito in maniera uniforme maggiore o uguale a 0 e minore di 1.

** **Nota**
Per poter utilizzare il computer è necessario trasformare il codice identificativo dell'alunno da alfabetico a numerico.

Ad esempio, con un foglio elettronico la sintassi da utilizzare per l'estrazione del campione è la seguente:

```
=INT(CASUALE()*6)+1   = CERCA.VERT(A56;$B$1:$C$7;2)
```

La prima formula consente la generazione casuale degli elementi del campione, la seconda consente di estrarre dalla tabella dei voti il voto corrispondente all'alunno estratto.

Di ogni campione viene calcolata la media e delle medie trovate dagli studenti viene costruita e analizzata la distribuzione.

La Tabella 4 mostra la distribuzione delle medie campionarie ottenute per simulazione di un campionamento di numerosità 2 da parte dei 25 alunni di una classe.

Distribuzione della media campionaria dei voti per campioni di 2 elementi da una popolazione di 25

$\overline{x}_i$	n_i	$\overline{x}_i \cdot n_i$	$\overline{x}_i^2 \cdot n_i$
4	2	8	32
4,5	6	27	121,5
5	14	70	350
5,5	16	88	484
6	27	162	972
6,5	16	104	676
7	10	70	490
7,5	7	52,5	393,75
8	2	16	128
Tot	100	597,5	3647,25
$M(\overline{X})$	5,975		
$Var(\overline{X})$	0,771875		

Tabella 4

Il grafico di Figura 2 mostra la rappresentazione grafica della distribuzione ottenuta con la simulazione:

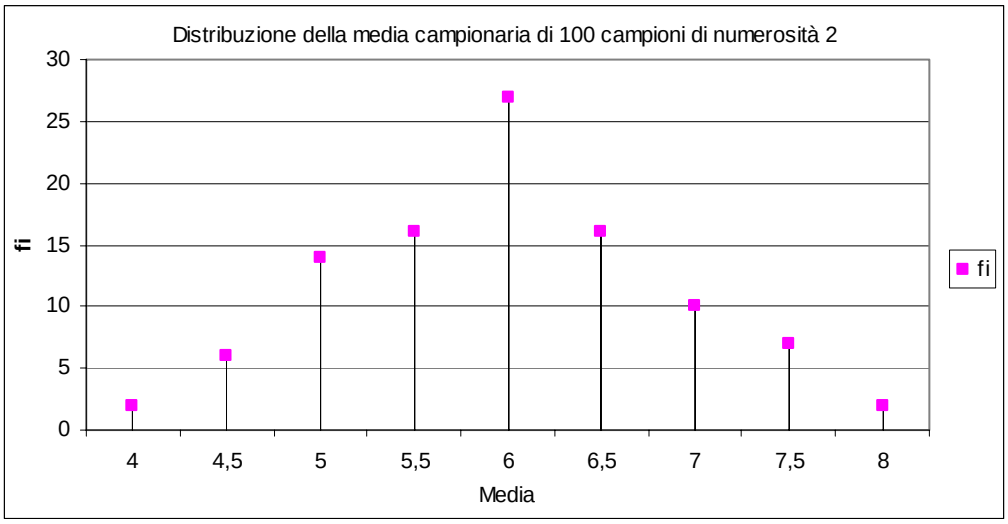


Figura 2

L'insegnante invita gli studenti a confrontare i valori ottenuti attraverso la loro esperienza di simulazione con i dati ricavati utilizzando tutti i possibili campioni appartenenti allo spazio campionario dell'esperimento casuale di numerosità 2 descritto precedentemente.

L'andamento del grafico in Figura 2 non è perfettamente simmetrico. La forma è simile al grafico di Figura 1? Il valore 6 rappresenta il valore più frequente (moda) in entrambe le distribuzioni?

Confrontando gli indici si nota che la media delle "medie campionarie" nella simulazione è vicina alla media della popolazione, mentre la sua varianza, pur rimanendo minore della varianza della popolazione, non conferma appieno la proprietà della varianza campionaria.
L'insegnante pone la domanda: perché non sono uguali? Dipende dalla quantità e dal tipo di campioni estratti? Siamo sicuri che tra i campioni estratti non ce ne siano di ripetuti? Siamo sicuri che tra i 100 campioni degli studenti ci siano tutti quelli dello spazio campionario?

Come ulteriore attività l'insegnante propone di aumentare la numerosità del campione passando da 2 elementi per campione a 3 elementi. Invita pertanto gli studenti a ripetere le operazioni già eseguite precedentemente su 106 campioni e ad analizzare la distribuzione delle medie campionarie ottenute, di cui si fornisce un esempio di distribuzione e il relativo grafico.

Distribuzione della media campionaria dei voti per campioni di 3 elementi da una popolazione di 6

$\overline{x}_i$	n_i	$\overline{x}_i \cdot n_i$	$\overline{x}_i^2 \cdot n_i$
4,33	1	4,333	18,778
4,67	6	28,004	130,704
5,00	9	45,012	225,120
5,34	10	53,353	284,658
5,67	12	68,032	385,696
6,00	18	108,060	648,720
6,34	16	101,397	642,589
6,67	14	93,399	623,094
7,01	9	63,048	441,672
7,34	8	58,715	430,927
7,67	3	23,020	176,640
	106	646,373	4008,597
$M(\overline{X})$	6,098		
$Var(\overline{X})$	0,633		

Tabella 5

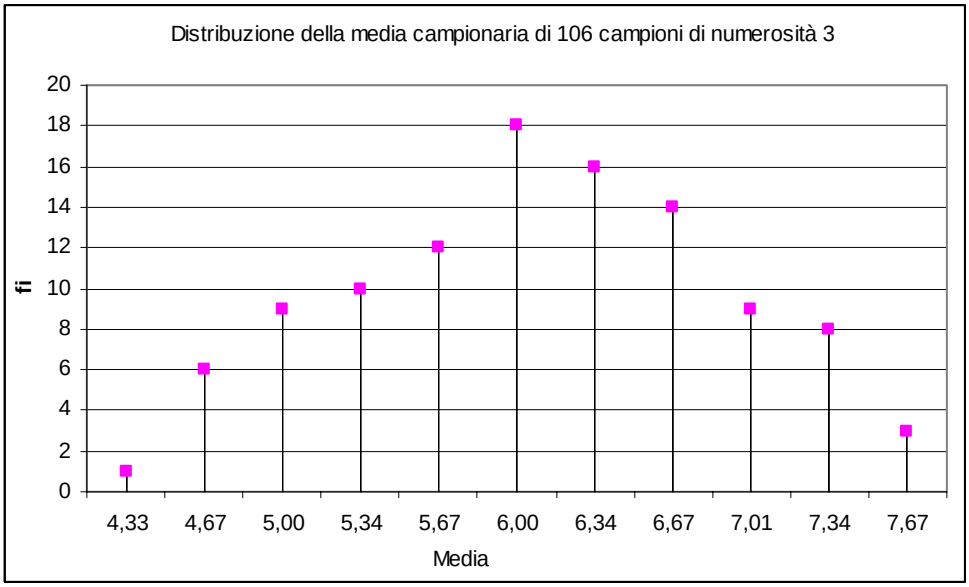


Figura 3

L’insegnante invita gli studenti ad analizzare i risultati ottenuti e a confrontarli con quelli ottenuti da campioni di numerosità 2. La distribuzione si mantiene simmetrica e a campana? Il valore della varianza è cambiato? E’ aumentato o diminuito? Che valore assume il rapporto tra la varianza della popolazione e la varianza campionaria?
L’insegnante chiede poi se è possibile aumentare ancora la numerosità del campione. Ad esempio: 6, 10, 30? Cosa succederà alla varianza della media campionaria?

Riferimenti bibliografici

Cicchitelli, G., (2001), *Probabilità e statistica*, Maggioli ed., Rimini.

Brusati, E., (2003), *Come si fanno i sondaggi*, *Induzioni*, n. 26, Istituti Editoriali e Poligrafici Internazionali, Pisa, p. 5 – 37.

Parpinel, F.; Provasi, C., (2004), *Elementi di probabilità e statistica per le scienze economiche,* Giapichelli ed., Torino.