Statistica sociale – 4

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

La metodologia delle indagini campionarie

Il campionamento - 1

- Le indagini campionarie non esauriscono la più vasta gamma delle indagini sociali
- Metodologie di indagine di natura più propriamente ***qualitativa***,
- come ad esempio le interviste non strutturate a testimoni privilegiati, i ***focus groups*** e così via.

Il campionamento - 2

- **Indagini** svolte con la raccolta di informazioni tramite uno **strumento di rilevazione** (prevalentemente) **strutturato** su un **campione** delle unità statistiche che costituiscono la popolazione sulla quale si vogliono inferire i risultati, campione scelto con criteri probabilistici per garantire la possibilità di effettuare l'**inferenza**.

Il campionamento - 3

- ricerche di mercato, *audience* dei media, sondaggi di opinione e così via;

- faremo riferimento, come *caso di studio* per i temi sociali, all'Indagine Multiscopo dell'Istat su "I cittadini e il tempo libero".

Le fasi delle indagini campionarie

L’individuazione della popolazione

- individuazione della popolazione che vogliamo studiare, ovvero la sua **collocazione temporale**, **geografica** e **demografica**.
- Nel caso di studio la popolazione italiana dai 3 anni in poi, nel periodo di svolgimento dell’indagine (2015).
- Tenere conto delle informazioni: liste anagrafiche, liste elettorali e collocazione sul territorio delle unità statistiche

  per individuare il campione.

Il tema della ricerca - 1

- Il secondo aspetto è l'individuazione dei **temi** sui quali si vogliono raccogliere le informazioni.

- Ciascun tema deve essere concettualmente definito, per la costruzione del questionario che permetta di rendere operativi i concetti individuati.

Il tema della ricerca - 2

- Prendiamo ad esempio il tema relativo alla **pratica sportiva** dei cittadini italiani, o meglio alla loro partecipazione ad attività sportive e/o fisico motorie.
- Se non viene individuata una definizione precisa di cosa sia un'attività sportiva, in senso lato di cosa sia "***sport***", qualunque risultato può essere interpretato in modo diverso, spesso contrastante.

Il lavoro sul campo

- Il terzo aspetto è quello delle modalità di svolgimento del lavoro sul campo, ovvero come **somministrare** il questionario, ovvero con:
  - intervista diretta (*face to face*),
  - intervista telefonica,
  - postale (posta tradizionale o via *e-mail*),
- con l'aiuto di un supporto informatico per registrare le informazioni (CATI, CAWI, CAPI), o con modalità miste fra quelle elencate.

La registrazione, elaborazione e analisi delle informazioni

- Una volta raccolti i questionari, i passi della ricerca saranno:
  - la codifica,
  - la registrazione (***input***) delle informazioni (se non già avvenuto come accade nelle indagini *Computer Aided*),
  - la loro elaborazione,
  - l'individuazione degli indici e degli indicatori da analizzare.

Gli *errori* - 1

- In ognuna di queste fasi è insito il rischio di ***errori***

- a parte quelli della **stima campionaria**, che sono statisticamente misurabili, tutti gli altri possono condurre a distorsioni nei risultati e a far fallire l'indagine.

Gli *errori* - 2

- le mancate risposte, ovvero unità statistiche che rifiutano di partecipare all'indagine o non rispondono a una o più delle domande del questionario,

- una imprecisa definizione del tema di ricerca, o una sua confusa e fuorviante traduzione in domande del questionario;

- errori nel processo di codifica e registrazione dell'informazione e così via.

La popolazione e il campione

Teoria dei campioni?

- In questo paragrafo non si vuole effettuare una trattazione rigorosa della Teoria dei campioni, ben più appropriatamente sviluppata nei corsi *ad hoc*: si vuole dare solo un'idea di base del campionamento, per capire la logica dei passi effettuati nel nostro caso di studio.

Quali informazioni?

- Le informazioni richieste possono essere riferite a un aspetto ***qualitativo*** (ad esempio se si pratica o meno uno sport): variabile dicotomica (“1”-“sì” e “0”-”no”);
- oppure possono far riferimento a un aspetto ***quantitativo***, ad esempio quante volte alla settimana ci si allena.

- Nel primo caso si inferirà la **proporzione** di coloro che rispondono “sì” e si parlerà di inferenza sugli ***attributi***, nel secondo si inferirà il **numero medio** di allenamenti a settimana e si parlerà di inferenza sulle ***variabili***.

Qualitativo vs. Quantitativo - 1

- Gli aspetti tecnici nei due casi coincidono:
- ogni variabile quantitativa può essere trasformata in qualitativa (ad esempio l'età, misurata nel continuum, può essere aggregata in fasce: 3-12 anni, i bambini; 13-19 anni, gli adolescenti; 20-34 anni, i giovani; e così via).
- Ogni categoria così individuata può essere trattata come variabile dicotomica: "1" e "0".

Qualitativo vs. Quantitativo - 2

- D'altra parte la quantificazione in "0" e "1" di ciascuno degli attributi permette di calcolare misure statistiche di sintesi, quali **media aritmetica** e **scarto quadratico medio**,

- anche in questo caso, quindi, le procedure di inferenza utilizzano le stesse leggi (**distribuzione normale delle medie campionarie**, **teorema del limite centrale** e così via) sia nella stima per gli attributi che in quella per le variabili.

Statistiche e parametri

- Il valore calcolato tra le unità campionarie è definito come la ***statistica campionaria***, che utilizziamo per stimare il ***parametro*** incognito nella popolazione:

- così se nel campione il 24% delle unità statistiche pratica uno o più sport, questa statistica ci servirà per stimare quanti italiani, nella fascia d'età considerata, praticano uno o più sport.

Stima puntuale

- Utilizzando la statistica per stimare direttamente il parametro si ha una ***stima puntuale***, molto semplice ed efficace (gli sportivi nella popolazione sono il 24%),

- ma non è possibile in questo caso dare una misura accurata e ragionevole dell'errore che si può commettere:

  se il parametro fosse 24,1% o 40% la nostra stima sarebbe comunque errata!

Intervallo di confidenza

- Le leggi dell'inferenza ci suggeriscono allora di effettuare una stima per ***intervalli di confidenza*** (di fiducia):

- la percentuale di sportivi nella popolazione sarà compresa fra 24% meno un ε% (che si cerca di minimizzare) e 24% più ε%;

- l'età media dei praticanti la pallacanestro sarà compresa fra 18 anni meno ε anni e 18 anni più ε anni.

Test di ipotesi

- Un'ulteriore possibile miglioramento del processo di inferenza è nell'effettuare un ***test di ipotesi***;
- si formula un'ipotesi sul parametro da stimare e si verifica se la statistica campionaria è compatibile con questo valore:
  - ad esempio, se essa cade nell'intervallo di confidenza al 95% costruito intorno al parametro ipotizzato, si afferma che non si hanno sufficienti informazioni per rifiutare l'ipotesi; altrimenti questa viene rifiutata.

La distribuzione campionaria della media

- Concentriamo quindi l'attenzione sulla *statistica* più comunemente stimata: la media aritmetica. L'inferenza statistica si basa su due distinte fasi, una teorica la fase **deduttiva** e una pratica la fase **induttiva**.
- Nella fase deduttiva noi abbiamo tutte le informazioni sulla popolazione, il suo ammontare (**N**) e il valore dei parametri da stimare ($\boldsymbol{\mu_x}$ e $\boldsymbol{\sigma_x}$), e ci possiamo chiedere: prendendo un campione di dimensione **n**, estratto con reintroduzione da questa popolazione, quale valore potrà assumere la media aritmetica del campione $\mathbf{M_x}$?

....

- Possiamo rispondere a questa domanda se conosciamo la distribuzione campionaria della media aritmetica: questa infatti assume valori diversi da campione a campione, a seconda delle unità che in esso sono inserite.

....

- La media campionaria è quindi una variabile che assume il più delle volte valori uguali o simili a quella della popolazione e meno frequentemente valori distanti da quella, più ci si allontana più la frequenza si riduce: siamo quindi in presenza di una tendenza alla distribuzione normale e si può dimostrare che la sua media aritmetica (la media delle medie campionarie) è uguale proprio a $\boldsymbol{\mu_x}$, mentre il suo scarto quadratico medio (anche detto ***errore standard*** – $\boldsymbol{\sigma_M}$) è uguale a $\boldsymbol{\sigma_x / \sqrt{n}}$.

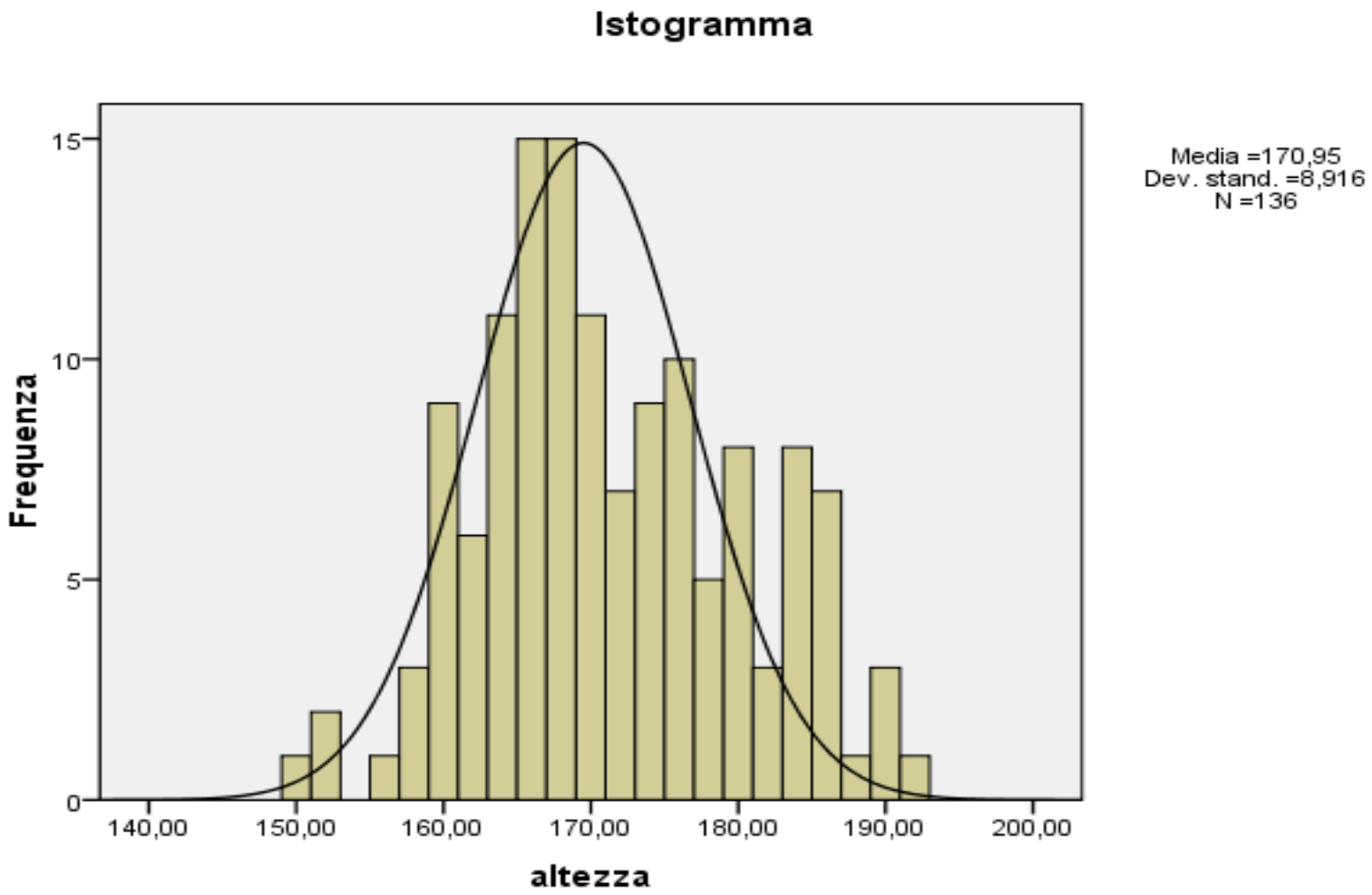
....

- Esiste un **teorema,** chiamato del **limite centrale**, che può essere rozzamente così sintetizzato:
- “Per campioni casuali di elevata ampiezza, la distribuzione campionaria della media campionaria tende alla distribuzione normale anche quando la variabile di cui si vuol calcolare la media non ha una distribuzione normale”, ossia diventa normale per $n \to \infty$, ma già è molto vicina per $n \geq 50$.

....

- Il fatto che l'errore standard sia uguale allo scarto quadratico medio della variabile studiata **diviso** per la radice della dimensione del campione (n), ci indica che la dispersione della variabile media campionaria è molto più piccola di quella di partenza (nei campioni i valori più alti della media si compensano con quelli più bassi) e questo è tanto più probabile quanto più alta è la numerosità del campione: se ad esempio n=100, allora σ_M è 10 volte più piccolo di σ_x.

La distribuzione normale - 1



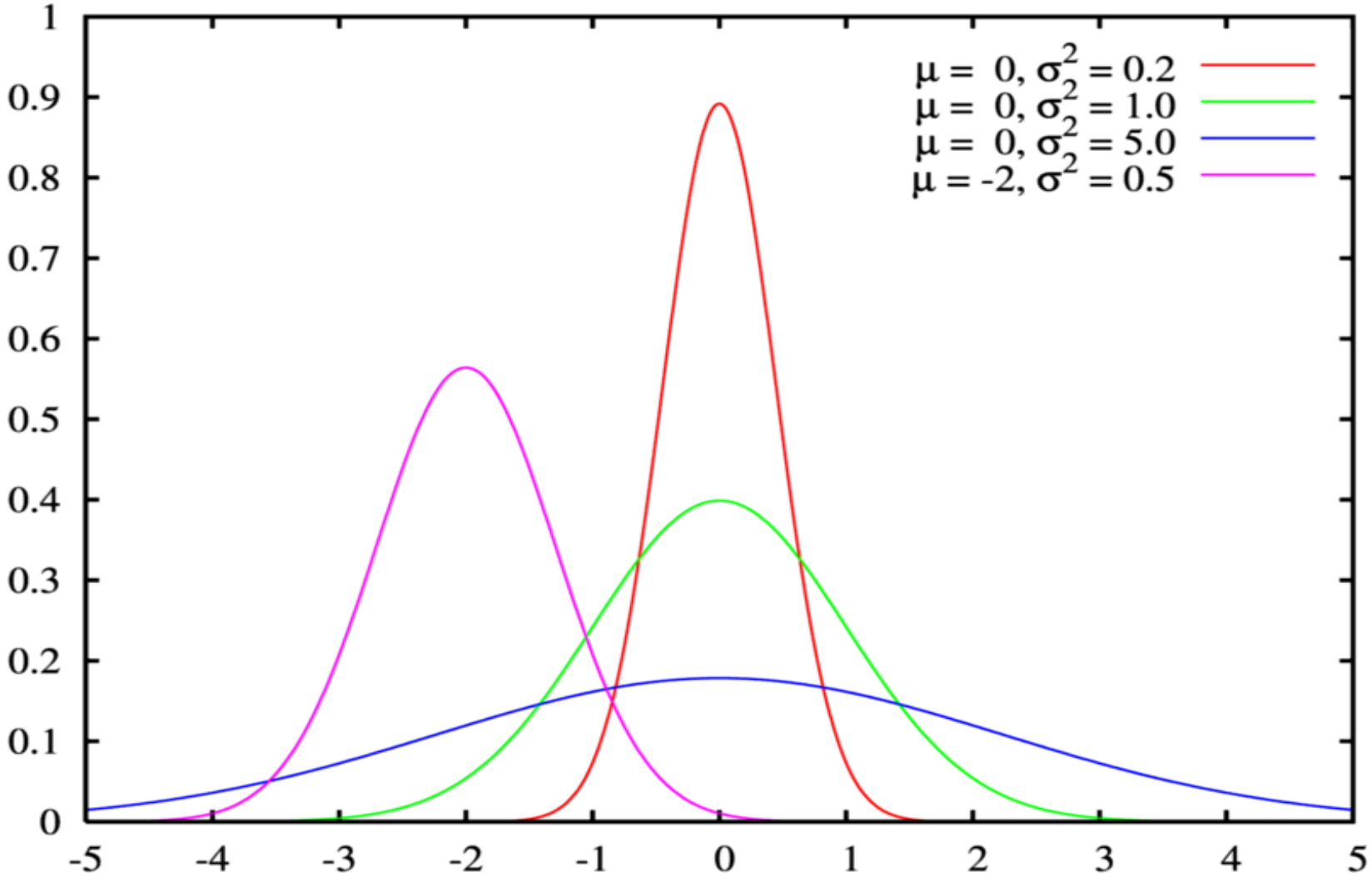
La distribuzione normale - 2

- Come nella rappresentazione ad istogramma, l'asse delle ascisse è riservato ai possibili risultati, quello delle ordinate alle frequenze (assolute o percentuali).
- Poiché siamo nel *continuo* non ha senso considerare i segmenti corrispondenti ai singoli punti: si ragiona in termini di aree.
- L'area complessiva al di sotto della curva corrisponde al 100% dei casi.

La distribuzione normale - 3

- Esistono infinite curve normali (∞^2), che variano al variare del loro punto centrale (media) e della dispersione (sqm).

La distribuzione normale – 3 bis



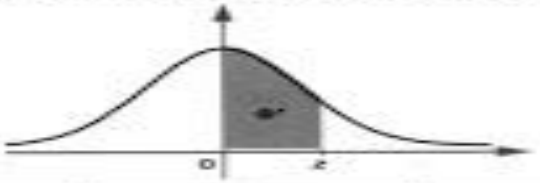
La distribuzione normale - 4

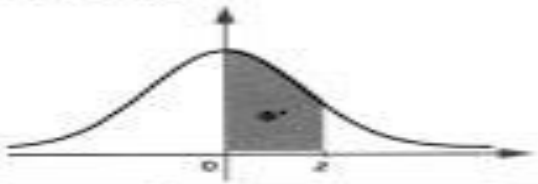
- Esiste, però,una sola **curva normale standardizzata**, ottenuta standardizzando la variabile studiata: in questo caso l'area corrispondente a ciascun valore, o coppia di valori, è tabulata (cfr. Tavole della Curva normale standardizzata).

Distribuzione normale $\phi^*(z)$

Area sottesa alla curva normale standardizzata da 0 a z
- test a due code: $z_{\alpha} = 1,96$; $z_{\alpha} = 2,58$
- test a una coda: $z_{\alpha} = 1,65$; $z_{\alpha} = 2,33$

La cifra in testa alla colonna rappresenta la seconda cifra decimale di z





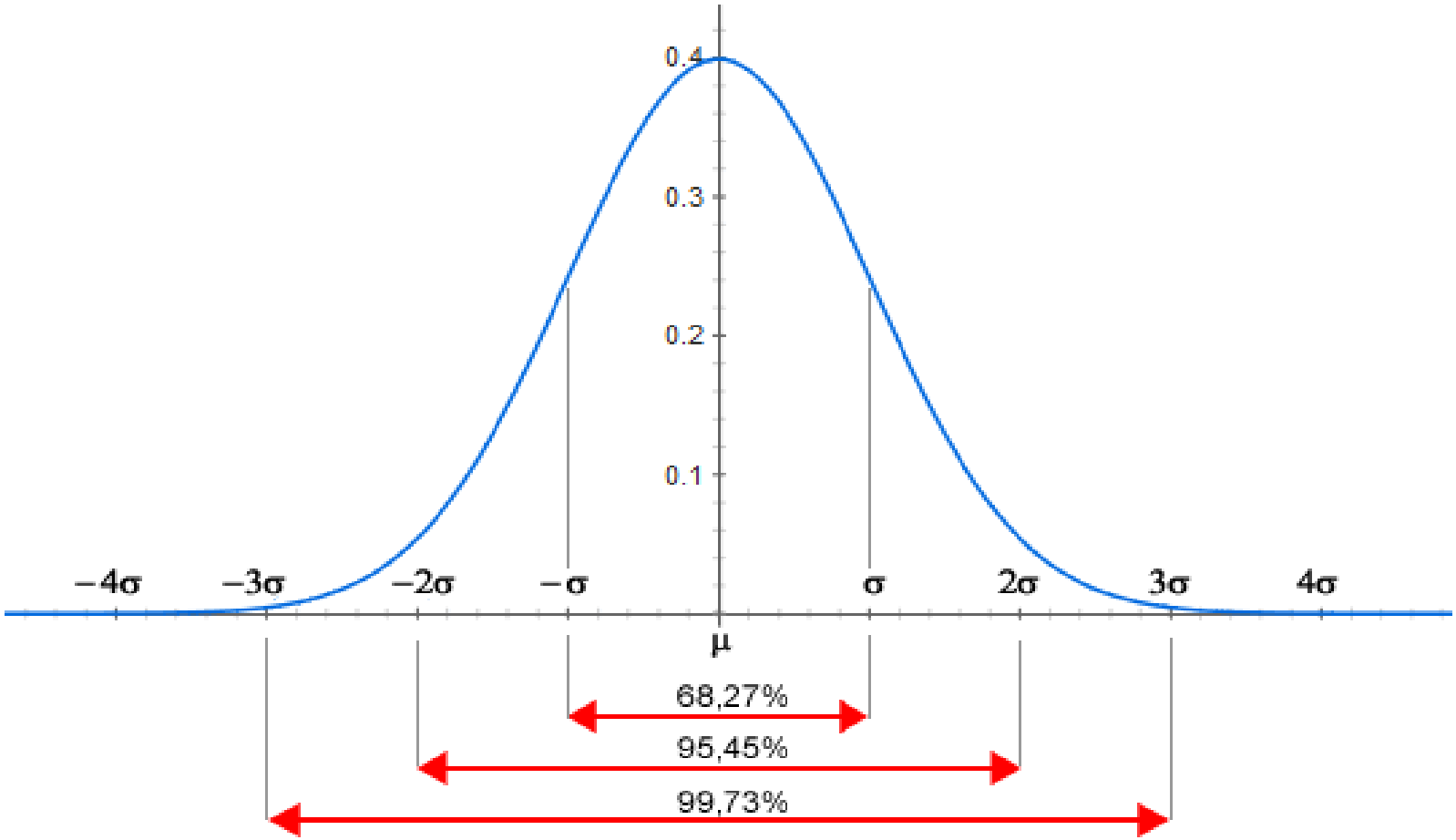
z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0754
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1809	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3706	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4811
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Esercizio

- Il tasso di glicemia nel sangue della popolazione ultrasessantenne è distribuito normalmente con un valore medio di 150 e uno scarto quadratico medio di 12.
- Qual è la percentuale di anziani che ha un tasso inferiore a 174?
- Qual è la percentuale di anziani che ha un tasso superiore a 138?
- Qual è la percentuale di anziani che ha un tasso compreso fra 145 e 155?
- Qual è il tasso che solo il 5% di anziani, malati di diabete, supera?

La distribuzione normale - 5

- Si può osservare come i valori sull'asse delle ascisse che corrispondono ai due *punti di flesso* della curva sono, rispettivamente la media meno lo sqm (a sinistra) e la media più lo sqm (a destra).
- L'area che sottosta alla curva fra questi due valori è pari al 68,26%.
- L'area che sottosta alla curva fra media meno 1,96 volte lo sqm e media più 1,96 volte lo sqm è pari al 95%, lasciando così nelle due code esterne il 5%!


0.4
0.3
0.2
0.1
−4σ
−3σ
−2σ
−σ
μ
σ
2σ
3σ
4σ
68,27%
95,45%
99,73%
WWW.OKPEDIA.IT

La distribuzione campionaria della media, esempio

- Cominciamo con un esempio molto semplice e registriamo la variabile “numero di televisori” in una popolazione di 5 famiglie (**N**=5); i valori sono: 0,1,2,3,4. $\boldsymbol{\mu_x}$ è facilmente calcolabile: è 2; $\boldsymbol{\sigma_x}$ è $\sqrt{2}$.
- Consideriamo ora tutti i campioni che è possibile estrarre da questa popolazione che abbiano dimensione pari a 2 (**n**=2).

Esempio

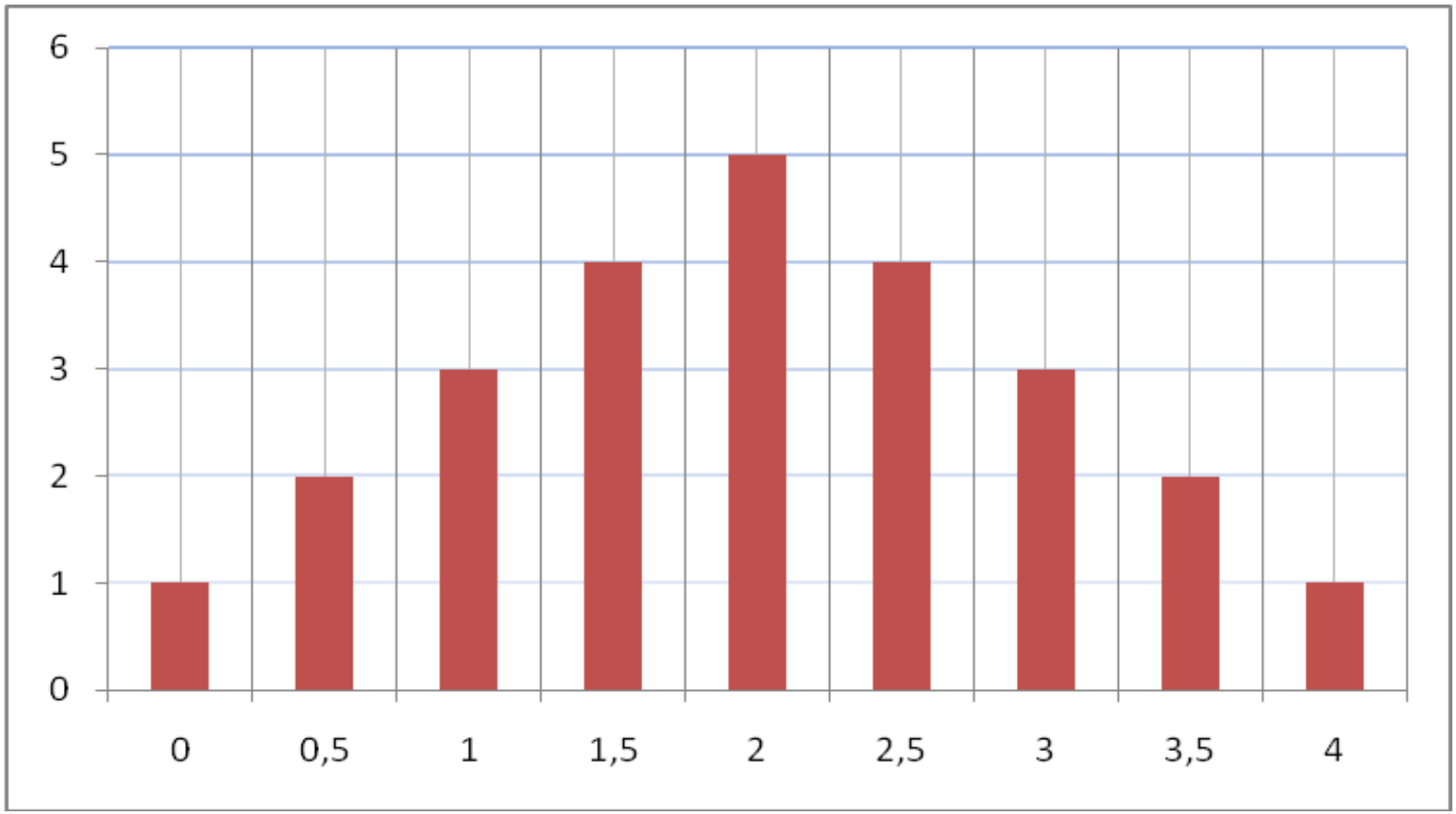
Ricordando le regole del calcolo combinatorio, troviamo che sono 25 ($\mathbf{N^n}$), per ognuno è calcolata la media aritmetica:

(0,0) **0**	(0,1) **0,5**	(0,2) **1**	(0,3) **1,5**	(0,4) **2**
(1,0) **0,5**	(1,1) **1**	(1,2) **1,5**	(1,3) **2**	(1,4) **2,5**
(2,0) **1**	(2,1) **1,5**	(2,2) **2**	(2,3) **2,5**	(2,4) **3**
(3,0) **1,5**	(3,1) **2**	(3,2) **2,5**	(3,3) **3**	(3,4) **3,5**
(4,0) **2**	(4,1) **2,5**	(4,2) **3**	(4,3) **3,5**	(4,4) **4**

....

- Se calcoliamo ora la media della variabile “media campionaria” vediamo che essa è ancora 2, mentre il suo s.q.m. è 1, ossia $\sigma_x/\sqrt{n}$ ($\sqrt{2}/\sqrt{2}$).
- Sono così verificate empiricamente le relazioni enunciate; se poi rappresentiamo graficamente (figura seguente) la distribuzione delle medie campionarie possiamo osservare come, pur non essendo normale, vi sia un rozzo approccio alle caratteristiche di questa distribuzione e questo per il valore di “n” più piccolo possibile!
- È prevedibile, quindi, che la normalità delle distribuzioni delle medie valga per n->∞.

Distribuzione di frequenza delle medie campionarie (con reintroduzione)



....

- Ma un campionamento con reintroduzione è poco realistico: perché dovrei campionare due volte la stessa famiglia? Perché dovrei considerare campioni diversi quello con una prima unità che ha 1 televisore e con una seconda che ha 2 televisori e quello con 2 e 1? Con un campionamento in blocco restano solo i campioni sottolineati:
-

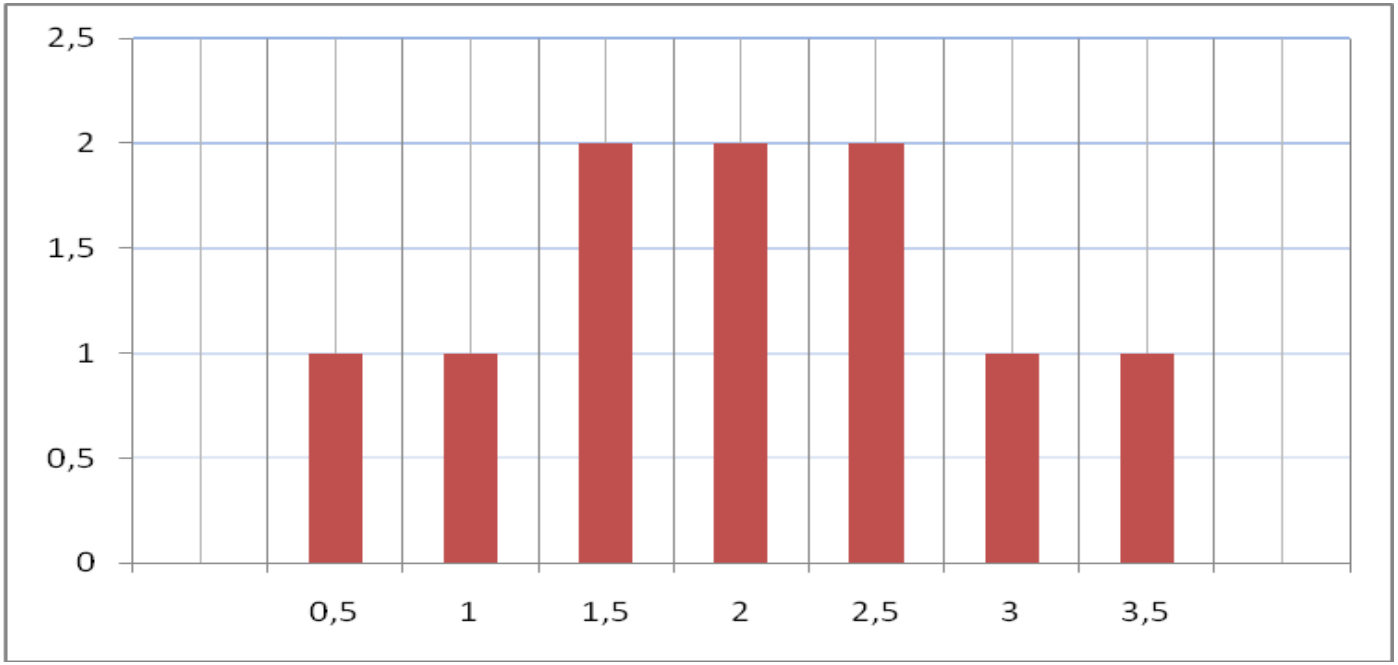
....

(0,0) 0	(0,1) 0,5	(0,2) 1	(0,3) 1,5	(0,4) 2
(1,0) 0,5	(1,1) 1	(1,2) 1,5	(1,3) 2	(1,4) 2,5
(2,0) 1	(2,1) 1,5	(2,2) 2	(2,3) 2,5	(2,4) 3
(3,0) 1,5	(3,1) 2	(3,2) 2,5	(3,3) 3	(3,4) 3,5
(4,0) 2	(4,1) 2,5	(4,2) 3	(4,3) 3,5	(4,4) 4

....

- Nella Figura vi è la sintesi dei risultati riportata come distribuzione di frequenza: la media è sempre 2, lo s.q.m. è più piccolo, perché non ci sono i valori estremi.

Distribuzione di frequenza delle medie campionarie (in blocco)



....

- Questi esempi sono molto interessanti perché ci evidenziano la possibilità di usare le Tavole della distribuzione normale standardizzata per ***misurare l'errore campionario***.
- Già abbiamo visto come si possono trovare le aree che sottostanno la distribuzione normale e come dalle aree si può tornare a calcolare i valori delle distribuzioni.
- Ora ci concentriamo su un caso particolare: ricorderete che tra i valori $\mu_x - 1{,}96\sigma_M$ e $\mu_x + 1{,}96\sigma_M$ cadono il 95% delle medie campionarie.
- Ciò vuol dire che 95 volte su 100 la media campionaria starà nell'intervallo, solo 5 volte ne starà fuori!

L'intervallo di confidenza

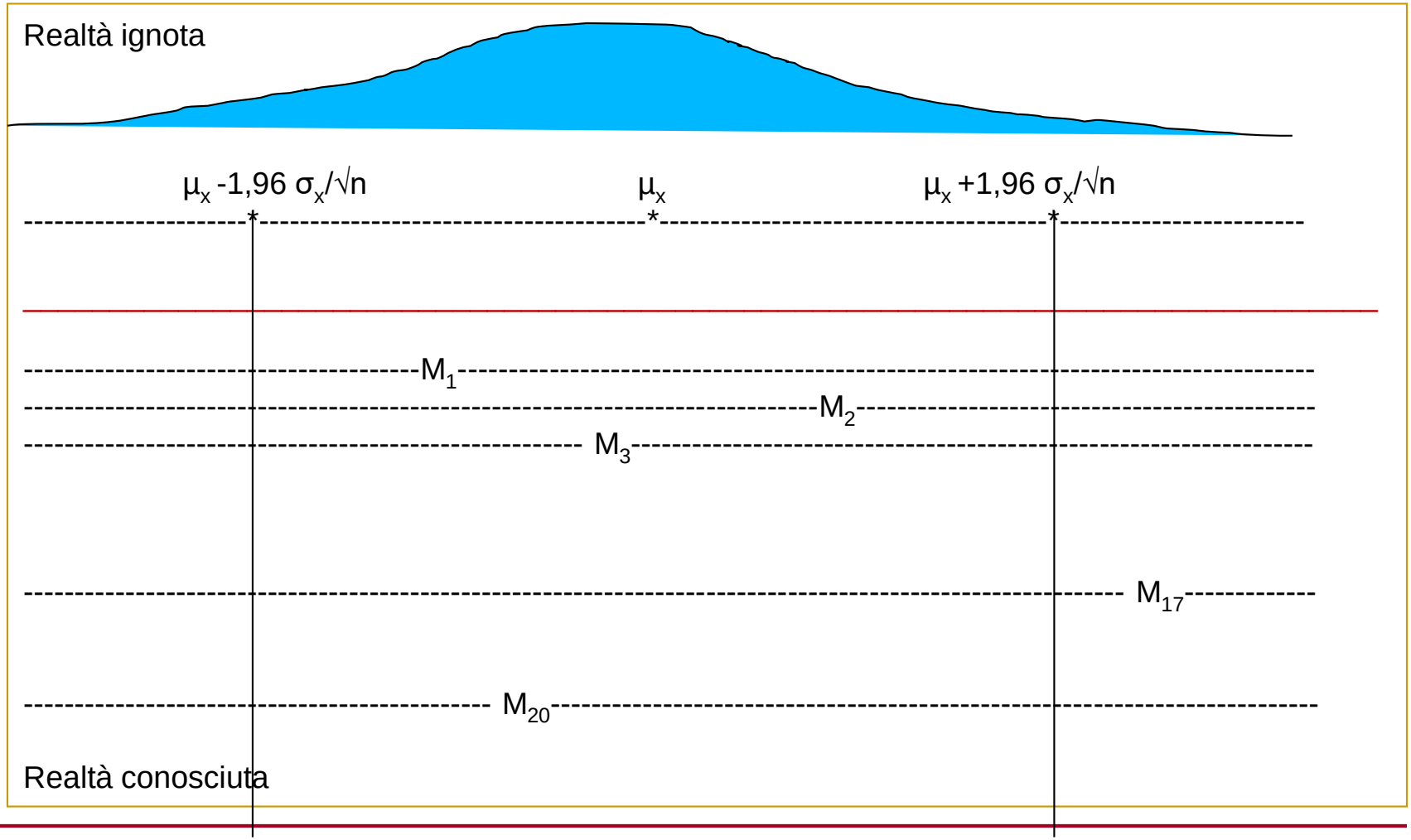
- Questo risultato è importante teoricamente (è il risultato della **deduzione**) ma del tutto inutile praticamente, perché noi non conosciamo né μ_x, né σ_x, sappiamo solo che la distribuzione delle medie campionarie è normale e che vale l'intervallo precedente, che chiamiamo **intervallo di confidenza** (o, meglio, ***intervallo di fiducia***). Proviamo allora a invertire le relazioni nell'intervallo (a):
  - Prop. ($\mu_x - \varepsilon \leq M_x \leq \mu_x + \varepsilon$) = 95% (a)
  - Prop. ($M_x - \varepsilon \leq \mu_x \leq M_x + \varepsilon$) = 95% (b)

Ricordiamo che $\varepsilon = 1{,}96\ \sigma_x / \sqrt{n}$

L’induzione

- Passando da (a) a (b) siamo entrati nella fase dell’**induzione**: non c’è più un intervallo fisso al cui interno *fluttua* il 95% delle medie campionarie, l’intervallo è variabile, a seconda della media campionaria che abbiamo calcolato e nel 95% dei casi la media della popolazione vi cade dentro.

Schema di passaggio deduzione vs. induzione



L'induzione

- Nella figura precedente si può vedere la situazione: la vera media della popolazione è $\boldsymbol{\mu_x}$ e l'intervallo è costituito da un suo intorno;
- nella fase della deduzione li conosciamo e possiamo vedere che su venti medie campionarie (per esempio ne sono riportate solo cinque) solo M_{17} cade fuori dall'intervallo;
- nella fase dell'induzione noi non conosciamo la parte alta del grafico, anzi abbiamo solo una delle venti medie campionarie: se questa fosse M_1 o M_2 o M_{20} si può vedere come $\boldsymbol{\mu_x}$ cada nell'intervallo; ma se conosciamo solo M_{17} allora $\boldsymbol{\mu_x}$ ne è fuori e la nostra stima è errata.
- Ma questo capita solo una volta su venti, anche se siamo senza difesa quando ci *dice particolarmente male*!!!

Commenti

- Si possono fare alcune osservazioni sull'intervallo considerato, che si basa sui valori **1,96**, $\boldsymbol{\sigma_x}$ e $\sqrt{\mathbf{n}}$:
- **±1,96**
- È un valore della variabile normale standardizzata cui corrisponde un'area pari al 95% di tutta l'area sottesa alla curva; pertanto nella coda sinistra si lascia un 2,5% e così nella coda destra, per un totale di un 5% di casi in cui di fatto facciamo una stima errata.
- Se vogliamo ridurre l'errore, dobbiamo lasciare nelle code a sn e a dx una percentuale inferiore (ad esempio 0,5%, per avere un errore solo nell'1% dei casi), ma allora il valore da inserire nell'intervallo aumenta (nell'esempio diventa ±2,54) e l'intervallo aumenta, ossia la stima è meno precisa.

....

- Se ci accontentiamo di una stima più rischiosa il valore, ovviamente, diminuirà e di conseguenza l'intervallo sarà più piccolo e la stima più accurata.
- σ_x
- Rappresenta la variabilità rispetto alla media della variabile di partenza: più il fenomeno che stiamo studiando è disperso, più l'intervallo è grande e la stima meno precisa.
- Su questo valore non si può intervenire perché è un dato fisso, il problema è che non è quasi mai conosciuto: si può stimare, con informazioni disponibili a priori da parte del ricercatore, oppure tramite s_x calcolato nel campione (ricordate la divisione per n-1?).

....

- Quest'ultimo caso deve essere affrontato a parte, perché la distribuzione della media campionaria quando si utilizza lo s.q.m. del campione per stimare quello della popolazione non è più normale, ma diventa una *t di Student*.
- **n**
- È la dimensione del campione: è un valore sul quale si può *giocare* molto per avere una stima più accurata. Esso è, infatti, posto al denominatore e più è grande, più l'intervallo si riduce, anche se non in maniera proporzionale, perché ne consideriamo la radice quadrata.