La v.a. normale standardizzata

La distribuzione normale standardizzata

- La distribuzione normale è difficilmente trattabile dal punto di vista calcolatorio, a causa dei suoi due parametri, μ e σ^2 .
- Il ricorso alla **"distribuzione normale standardizzata"** permette invece di individuare facilmente le probabilità relative agli intervalli di valori, utilizzando opportune *tavole statistiche*.

La distribuzione normale standardizzata

- **La distribuzione normale standardizzata** (detta "Z") si ottiene mediante una trasformazione della variabile X, di questo tipo ("punteggi z"):

$$z = \frac{(X - \mu)}{\sigma}$$

e pertanto

$$z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

La distribuzione normale standardizzata

La **standardizzazione** è una **trasformazione dei dati** che consiste nel:

- rendere la media **nulla** (μ = 0), dato che ad ogni valore della variabile originaria viene sottratta la media della variabile stessa;
- assumere la deviazione standard σ quale **unità di misura** (σ = 1) della nuova variabile, dato che ogni valore viene diviso per σ.

La distribuzione **normale standardizzata** viene indicata con **N(0,1)**.

I valori della Z sono **tabulati**: tra qualche diapositiva vedremo la tavola della Z.

La distribuzione Normale standardizzata

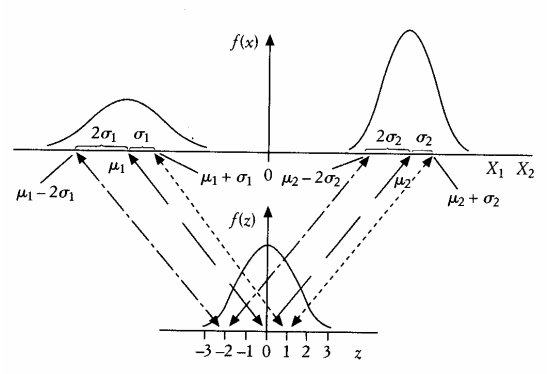
- La funzione di densità di probabilità della distribuzione normale standardizzata, *f* (z), assume la forma:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$\left(-\infty < z < +\infty\right)$$

Osservazione: questa funzione non contiene più i parametri.

La distribuzione normale standardizzata



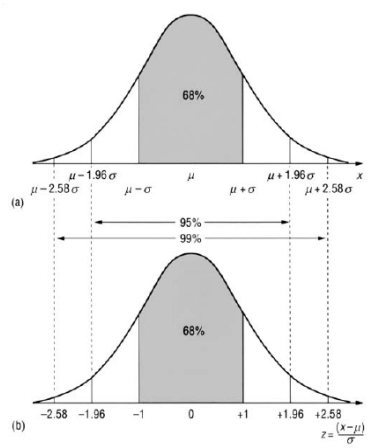
La v.a. normale standardizzata ha

MEDIA=0 e

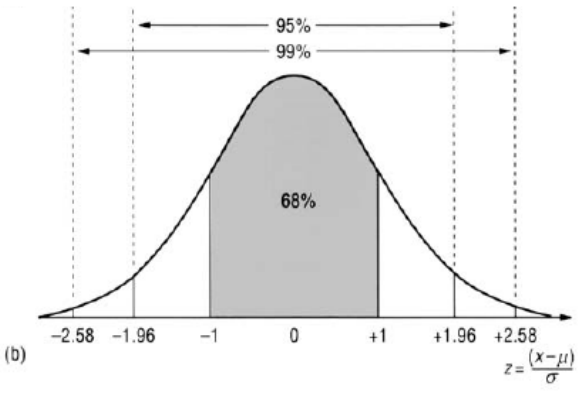
DEVIAZIONE STANDARD=1, per cui è rappresentata da UNA SOLA CURVA, mentre la distribuzione normale generale è rappresentata da infinite curve, che variano a seconda dei valori di μ e σ.

Aree sottese dalla curva normale generale

- La probabilità che un valore estratto casualmente da una v.a. $N(\mu,\sigma^2)$ sia compreso nell'intervallo $(\mu-\sigma, \mu+\sigma)$ è pari al **68%**;
- Il 95% dei valori assunti da una distribuzione Normale cadono nell'intervallo
- $(\mu-\mathbf{1,96}\sigma, \mu+\mathbf{1,96}\sigma)$;
- Il 99%, invece, nell'intervallo
- $(\mu-\mathbf{2,58}\sigma, \mu+\mathbf{2,58}\sigma)$



Aree sottese dalla curva normale standardizzata



Aree sottese dalla curva normale standardizzata

- La distribuzione normale standardizzata è importante perché le probabilità corrispondenti alle aree sottese dalla curva normale possono essere calcolate.
  Queste probabilità vengono riportate in apposite tavole.

  - In questo modo è possibile evitare il ricorso a complessi calcoli integrali per trovare le probabilità che una v.a. X assuma valori compresi all'interno di determinati intervalli.

Aree sottese dalla curva normale standardizzata

- È noto che il 68,26% dell'area totale è compreso tra ±1 *deviazioni standard* attorno alla media, cioè a ±1 *punti z* dalla media; mentre il 95,44% è racchiuso tra ±2 *deviazioni standard* attorno alla media: quindi a ±2 *punti z* dalla media.

Aree sottese dalla curva normale standardizzata

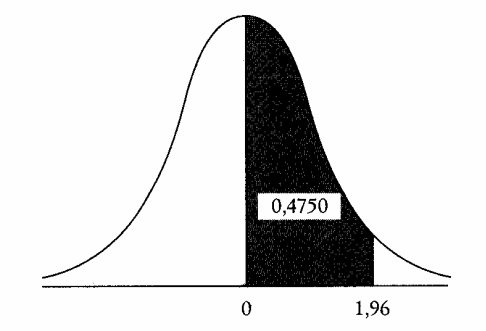
In virtù della proprietà di simmetria della distribuzione normale, le tavole riportano soltanto i valori dell'area compresa fra lo zero e l'ascissa +X, poiché, per la simmetria, l'area sottesa dall'altra metà della curva è ovviamente uguale.

Aree sottese dalla curva normale standardizzata

- Osservando la tavola, si troveranno i punti z nella colonna di sinistra con una cifra decimale; la seconda cifra decimale è posta nella prima riga in alto della stessa tavola.

In termini pratici …

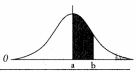
Supponiamo di voler conoscere l'area compresa tra le ascisse pari, rispettivamente, a z=0 e z=1,96.



In termini pratici …

- Osservando la colonna dei punti z, si deve scendere fino a trovare z=1,9, e poi rimanere nella stessa riga fino a trovarsi in quella indicata con 6.
- Il punteggio che si trova in quel punto indica la porzione di area compresa tra i due valori di z: 0,4750. Poiché l'area totale sottesa dalla curva nella sua parte positiva è pari a 0,500, l'area che si trova alla destra del valore z =1,96 sarà data da:
- 0,5000-0,4750=0,0250.

TAV. B. *Aree della distribuzione normale standard tra a = 0 e b > 0*



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	8.032	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4351	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

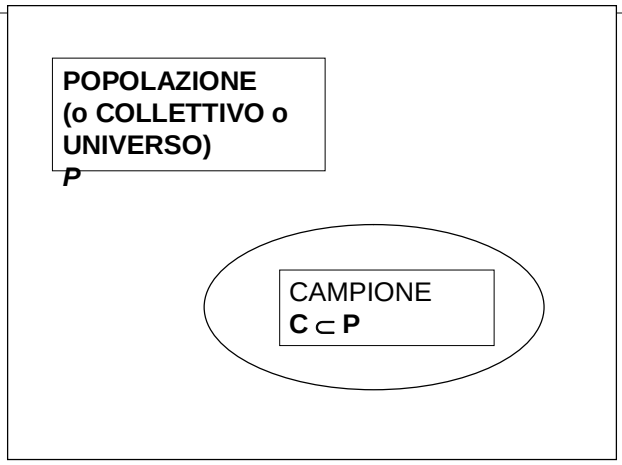
Inferenza statistica

Dalla popolazione al campione: l'inferenza statistica

- Il reperimento dei DATI STATISTICI attraverso una RILEVAZIONE è un'operazione che ha dei COSTI, sia in termini di TEMPO IMPIEGATO che in termini ECONOMICI.
- In molti ambiti scientifici, come in biologia e in medicina, si ha spesso a che fare con dati di origine sperimentale, per i quali quello di cui si dispone **è sempre un campione**, visto che la popolazione di riferimento è virtualmente "infinita" (es. un campione di animali da laboratorio rappresenta idealmente **tutti** gli esemplari di quella specie di animali da laboratorio).

Dalla popolazione al campione

- Nei primi anni del '900, integrandosi con alcuni risultati fondamentali del calcolo delle probabilità, la statistica ha cominciato ad interessarsi ai CAMPIONI.
- Un CAMPIONE è un SOTTOINSIEME del collettivo (popolazione) rispetto al quale si sta studiando un certo fenomeno: di solito, è un sottoinsieme di numerosità MOLTO INFERIORE a quella del collettivo di partenza.
- La TEORIA DELL'**INFERENZA STATISTICA** studia le tecniche per ricavare informazioni attendibili dai dati campionari.



- Le TECNICHE CAMPIONARIE (TEORIA DEI CAMPIONI) permettono di OTTIMIZZARE i criteri di ESTRAZIONE DEL CAMPIONE (il cosiddetto DISEGNO DI CAMPIONAMENTO), in maniera tale da **RICAVARE DAL CAMPIONE PRESSOCHE' LE STESSE INFORMAZIONI CHE SI SAREBBERO RICAVATE DISPONENDO DELL'INTERO COLLETTIVO**.

IN STATISTICA:

- ***RAPPRESENTATIVITÀ*** **DEL CAMPIONE**

=

- **ESTRAZIONE CASUALE DELLE UNITA' DEL CAMPIONE**

Statistica e calcolo delle probabilità

- Le tecniche di inferenza statistica si basano tutte sulla "somiglianza" (che in termini tecnici si chiama "verosimiglianza") del campione rispetto alla popolazione da cui è stato estratto; estratto, non dobbiamo mai dimenticarlo, con criteri rigorosamente **casuali**.
- L'inferenza statistica è resa possibile dalla conoscenza delle leggi fondamentali del calcolo delle probabilità.

Variabili aleatorie e distribuzioni teoriche di probabilità

- In statistica descrittiva abbiamo visto cosa intendiamo per **variabile** statistica. Spostandoci in campo probabilistico, se una variabile può assumere valori esclusivamente dovuti al caso (o, per essere più precisi, a un **esperimento aleatorio**), essa prende il nome di **variabile aleatoria**.
- Una **v.a.** è un numero **X** che assume un valore in **R** (asse dei numeri reali), determinato sulla base di un evento E, che si riferisce a un certo esperimento aleatorio. A ciascun valore assunto da X, legato all'evento E, si associa una probabilità P (P varia tra 0 e 1).
- Una **distribuzione di probabilità** è una funzione che sintetizza la relazione tra i valori di una variabile casuale X e la probabilità ad essi associata. Una distribuzione di probabilità descrive il comportamento della v.a. a cui è associata.

Variabili aleatorie e distribuzioni teoriche di probabilità

- La conoscenza della distribuzione di probabilità di una variabile aleatoria fornisce ai ricercatori uno strumento potente per descrivere una **popolazione**, dalla quale verranno estratti i campioni che, successivamente, saranno studiati.
- Molti fenomeni naturali e biologici si distribuiscono secondo una **distribuzione normale**.
- Una distribuzione di probabilità è solitamente rappresentata da una formula (*funzione di probabilità* se la v.a. è discreta, *funzione di densità* se la v.a. è continua)

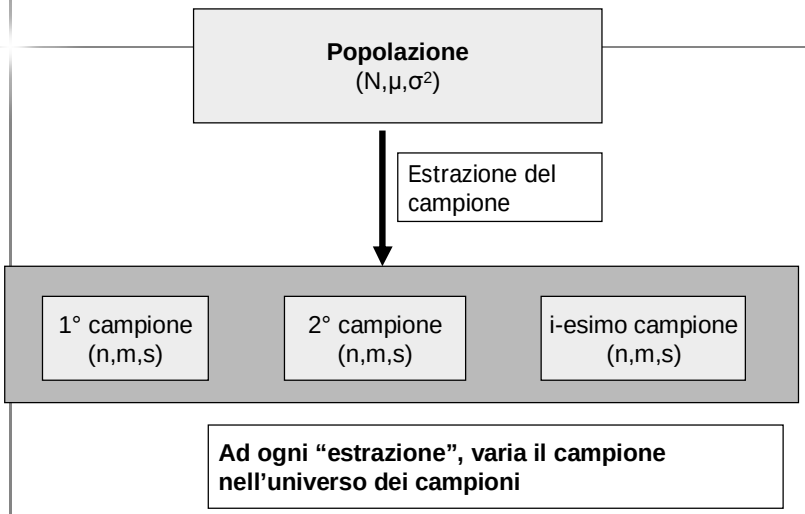
Variabili aleatorie e distribuzioni teoriche di probabilità

- La forma di una distribuzione può essere **simmetrica** rispetto al valore centrale, oppure vi può essere una coda più lunga da un lato piuttosto che dall'altro. Se la coda "lunga" è a destra (o, viceversa, a sinistra) la distribuzione avrà **asimmetrica positiva** (o, viceversa, **negativa**).
- Alcune distribuzioni teoriche di probabilità comunemente usate per descrivere i dati sono: la distribuzione Normale (o Gaussiana), la distribuzione Binomiale, la distribuzione di Poisson.
- Abbiamo già parlato ampiamente della v.a. **Normale** (o **Gaussiana**).

Dalla popolazione al campione: l'inferenza statistica

- Nelle indagini campionarie, l'obiettivo è quello di fare "**inferenza**" dal campione alla popolazione.
- In questo passaggio, poiché il campionamento è stato effettuato con criteri casuali, le "stime" che si ottengono sul campione, rispetto alla popolazione di partenza, sono per loro stessa natura affette da un errore, che si chiama errore di campionamento.
- Ad esempio, se io calcolo la media delle altezze di un campione di persone, per le quali so che la popolazione di partenza ha media 175 cm, non otterrò mai esattamente una media pari a 175 cm. Piuttosto, ripetendo infinite volte la stima su "infiniti" campioni, otterrò certamente una DISTRIBUZIONE DI MEDIE, che sarà dispersa attorno al vero valore della media, cioè 175 cm.

Popolazione e campione



La stima dei parametri di una popolazione

I parametri sono dei valori caratteristici della popolazione, come la media aritmetica, la probabilità del verificarsi di un evento, ecc.

Le stime sono effettuate in funzione delle osservazioni campionarie e, pertanto, dipendono dagli elementi del campione (media aritmetica del campione, frequenza di un certo evento nel campione, ecc.)

L'insieme di tutti i campioni estraibili casualmente da una popolazione è detto "spazio campionario". Se la popolazione è finita, si parla di "universo dei campioni".

Al variare del campione nell'universo campionario la stima assume valori diversi, valori dei quali sarà possibile costruire la distribuzione (distribuzione campionaria).

Stime e stimatori

- Per stimare un parametro (ad es. la media) della popolazione originaria si estrae un solo campione.
- Tutti i possibili campioni, virtualmente, sono estraibili, e sono pertanto possibili **diverse stime del parametro**, in numero corrispondente a quello dei possibili campioni.
- Si possono costruire pertanto le distribuzioni delle medie campionarie che, in termini probabilistici, costituiscono anch'esse delle variabile aleatorie, descrivibili da modelli discreti o continui.

Gli stimatori

- Lo **stimatore** è una variabile aleatoria definita nell'universo dei campioni; lo stimatore è una variabile aleatoria, che assume valori in ciascun campione compreso nell'universo dei campioni.
- Mentre una **stima** è una determinazione empirica (una "realizzazione") del corrispondente stimatore.
- Vediamo perché.

Stimatori e statistiche

- Supponiamo di voler stimare, nella popolazione, un certo parametro θ.
- A partire dai dati osservati sul campione è possibile calcolare una **statistica t**, cioè una certa **funzione dei dati del campione** utilizzata allo scopo di stimare il parametro θ incognito:
- $t = f(x_1, x_2, \ldots x_n)$

Stimatori e statistiche

- Una statistica è, pertanto, una funzione applicata all'insieme degli n dati del campione. Ognuno di questi dati, poiché i campioni variano nell'universo dei campioni, descrive a sua volta una variabile aleatoria.
- Queste variabili aleatorie, se il campionamento è perfettamente casuale, sono tra loro indipendenti e identicamente distribuite.

Stimatori e statistiche

- Da ciò consegue che la statistica si può vedere, in ultima analisi, come una particolare determinazione di una funzione di variabili aleatorie.
- Definiamo pertanto stimatore questa funzione, T_θ, funzione di *n* variabili aleatorie, utilizzata per stimare il parametro θ.
- $T_\theta = f\ (X_1, X_2, \ldots . X_n)$

Proprietà degli stimatori

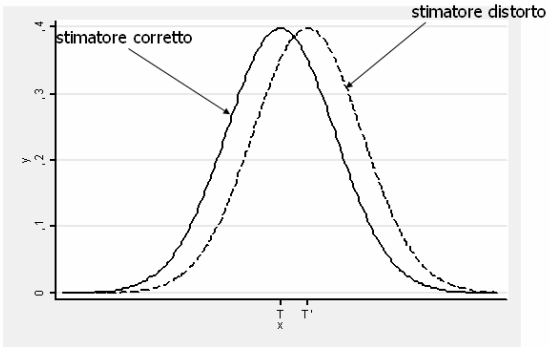
- Ad esempio, la **media campionaria** è uno **stimatore** del parametro "media" della popolazione;
- Come abbiamo visto, uno stimatore è una v.a., funzione dei valori del campione che, una volta calcolata, restituisce una **stima;**
- Uno stimatore può godere di alcune **proprietà** desiderabili.

Correttezza o non distorsione

Uno stimatore T si dice corretto, o non distorto, se il suo valore atteso coincide con il valore del parametro che intende stimare. Ad esempio la media campionaria è uno stimatore **corretto** della media della popolazione. Se, invece, il valore atteso non coincide con il parametro lo stimatore si dice **distorto**.

$$E(T_\theta) = \theta \quad \text{stimatore non distorto}$$
$$E(T_\theta) \neq \theta \quad \text{stimatore distorto}$$
$$\mathrm{B} = E(T_\theta) - \theta \quad \text{distorsione (bias)}$$

Distribuzione di uno stimatore



Esempi di stimatori corretti e distorti

- La media campionaria è uno stimatore corretto della media della popolazione: infatti, il suo valore atteso è proprio μ;
- La varianza campionaria è uno stimatore *distorto* (se *n* è piccolo) della varianza della popolazione

Varianza campionaria corretta

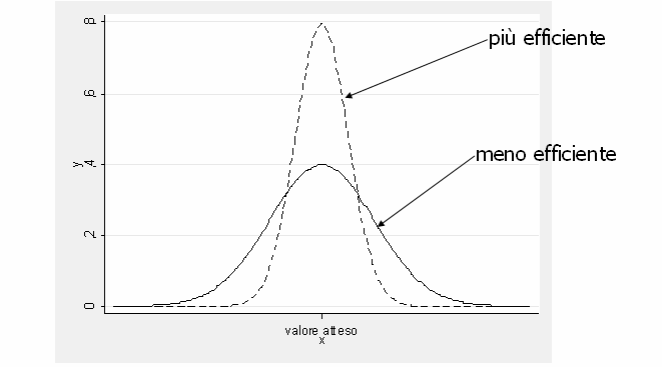
Si può dimostrare che la varianza campionaria **non** è uno stimatore corretto perché il suo valore atteso non coincide con σ^2. Si usa allora la sua "versione corretta" che si calcola nel seguente modo:

$$s^2 = \frac{\sum_{j=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

Efficienza

- Da uno stimatore ci aspettiamo non solo un valore medio uguale a quello della popolazione, ma anche che lo stimatore sia abbastanza “concentrato” attorno al valore medio.
- Se valutiamo due stimatori entrambi corretti, dobbiamo preferire quello con varianza minore perché è più **efficiente**, cioè riduce il margine di errore.

Confronto tra due stimatori entrambi corretti



Efficienza relativa

Quello di efficienza è un concetto sempre relativo. Dati due stimatori entrambi corretti, T_1 e T_2, si definisce "efficienza relativa" del primo rispetto al secondo il rapporto tra la varianza del secondo e la varianza del primo:

$$\text{Eff. relativa}\,(T_1, T_2) = \frac{Var(T_2)}{Var(T_1)}$$

Ad esempio, la mediana campionaria è meno efficiente della media campionaria perché ha una varianza più elevata. Il rapporto tra varianza della mediana e della media campionaria è circa 1.5. Dunque la media campionaria è preferibile come stimatore della media della popolazione.

Consistenza

- Se, all'aumentare della dimensione *n* del campione, il valore atteso di uno stimatore tende a concentrarsi intorno al vero valore del parametro, lo stimatore si dice **consistente**.

Metodi per il reperimento dello stimatore "migliore"

- Abbiamo visto finora quali sono le proprietà "auspicabili" per uno stimatore.
- Gli stimatori finalizzati alla stima di un parametro, però, possono essere vari: certamente non c'è un solo stimatore possibile.
- In questo corso, non ci addentriamo sui metodi con i quali è possibile reperire gli stimatori "ottimali" per i rispettivi parametri.
- Il metodo di gran lunga più utilizzato è il cosiddetto **"metodo della massima verosimiglianza"**.
- Si basa sul presupposto secondo cui il miglior stimatore possibile è quello che rende "più probabile" l'estrazione di un campione rispetto a una certa popolazione (la cui distribuzione è nota), dalla quale il campione viene estratto.
- Uno stimatore ottenuto con questo metodo si dice **stimatore di massima verosimiglianza** di un dato parametro, θ.

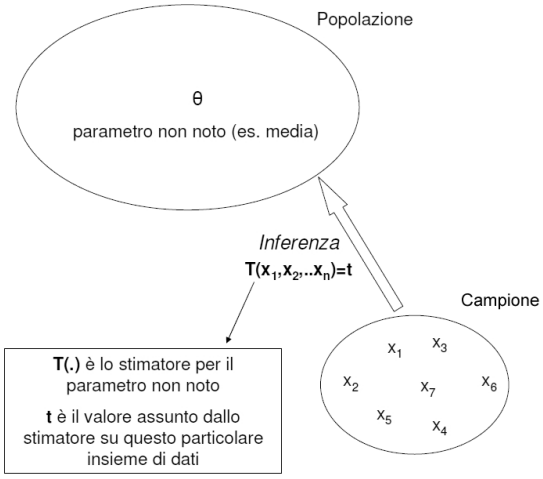
La variabilità campionaria

- Come abbiamo visto, una certa statistica, applicata al campione, il cui scopo è quello di stimare il corrispondente parametro incognito nella popolazione si dice **stimatore**.
- L'aspetto che va sottolineato è che uno stimatore, che è funzione di *n* variabili aleatorie, è esso stesso una variabile aleatoria.

La variabilità campionaria

- Infatti, il valore assunto dallo stimatore varia al variare dei dati sui quali si applica (varia al variare del campione).
- Si può definire la probabilità che lo stimatore assuma dei particolari valori. La probabilità sui possibili valori assunti dallo stimatore è indotta dalla **legge di probabilità** definita sullo spazio dei possibili campioni.
- La distribuzione di probabilità di una statistica è chiamata **distribuzione campionaria** della statistica.

Schematicamente



Distribuzione della media campionaria

- Sia *Y* una variabile aleatoria con media μ e varianza σ^2
- Sia dato un campione di numerosità *n* su cui si osserva la variabile ***Y***, che nel campione assumerà le determinazioni $Y_1, Y_2, \ldots, Y_n$
- La media campionaria è lo stimatore:

$$\overline{Y} = \frac{\sum_{j=1}^{n} Y_j}{n}$$

Media campionaria ed "errore standard"

- Si dimostra che la distribuzione dello stimatore "media campionaria" ha media μ (la stessa della popolazione da cui il campione è stato estratto) e varianza σ^2/n.
- Lo stimatore media campionaria, pertanto, è non distorto.
- La deviazione standard della distribuzione campionaria è chiamata **errore standard**.
- Il termine **errore standard** è utilizzato per distinguere la deviazione standard di una statistica campionaria da quella della popolazione da cui il campione è stato estratto.

Teorema centrale del limite

- Se un campionamento viene ripetuto “infinite” volte, per il **teorema centrale del limite,** la media campionaria tende a distribuirsi in modo normale, anche quando non lo è la popolazione da cui i campioni sono stati estratti.
- Se *n* è “sufficientemente grande” (>30), la forma della distribuzione campionaria delle medie è approssimativamente normale, indipendentemente dalla forma della distribuzione della popolazione di origine.

Teorema centrale del limite

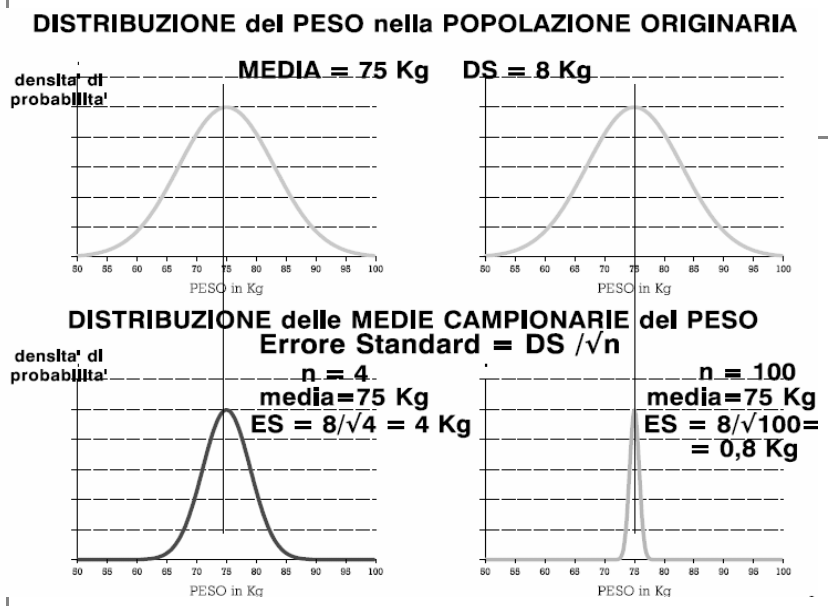
- Il valore medio dell’insieme di tutte le possibili medie campionarie è uguale alla media μ della popolazione di origine.
- La deviazione standard dell’insieme di tutte le possibili medie campionarie di campioni di numerosità *n*, detta **errore standard**, è funzione sia della deviazione standard della popolazione, sia della numerosità del campione:

$$ES(X) = \sigma_{\bar{X}} = \frac{\sigma_{POP}}{\sqrt{n}}$$

Errore standard <u>stimato</u> nel campione

Raramente conosceremo il valore di σ nella popolazione. Più spesso, dovremo **stimare** il valore di σ con il valore di s campionario (ottenuto con l'opportuno stimatore, che abbiamo visto in precedenza), e di conseguenza, calcolare il corrispondente valore dell'errore standard campionario.

$$ES(X)_{stimato} = s_{\bar{X}} = \frac{s_{campione}}{\sqrt{n}}$$



Un semplice esempio

- Una v.a., Y, assume valori discreti da 1 a 5, tutti con uguale probabilità (p=0,2=20%).
- La popolazione è formata da 5 unità.
- Si tratta di una una v.a. uniforme discreta.

Y	Freq.	Percent	Cum.
1	1	20.00	20.00
2	1	20.00	40.00
3	1	20.00	60.00
4	1	20.00	80.00
5	1	20.00	100.00
Total	5	100.00	

Un semplice esempio

- Supponiamo di estrarre un campione di n=2 unità dalla popolazione di N=5 unità.
- I momenti della v.a. media campionaria saranno quindi, secondo la regola appena vista:
- $E(Y_{medio}) = E(Y) = (1+2+3+4+5)/5 = 15/5 = 3$
- $Var(Y_{medio}) = s^2/n = [[(1-3)^2+(2-3)^2+(3-3)^2+ +(4-3)^2+(5-3)^2]/5]/2 = [[4 + 1 + 0 + 1 + 4]/5]/2 = =[10/5]/2 = 2/2 = 1$

Proviamo a verificarlo empiricamente

- Se estraiamo tutti i possibili campioni di 2 unità (che sono 5^2= **25**), otteniamo 25 possibili medie, qui schematizzate:

	1	2	3	4	5
1	1	1.5	2	2.5	3
2	1.5	2	2.5	3	3.5
3	2	2.5	3	3.5	4
4	2.5	3	3.5	4	4.5
5	3	3.5	4	4.5	5

Distribuzione empirica della media campionaria

- Se raccogliamo “in distribuzione” la tabella appena vista, otteniamo la distribuzione che segue:

$\bar{Y}$	Freq.	Percent	Cum.
1	1	4.00	4.00
1.5	2	8.00	12.00
2	3	12.00	24.00
2.5	4	16.00	40.00
3	5	20.00	60.00
3.5	4	16.00	76.00
4	3	12.00	88.00
4.5	2	8.00	96.00
5	1	4.00	100.00
Total	25	100.00	

Calcoliamo la media e la varianza della media campionaria

- $E(Y_{medio})$= (1+1,5*2+2*3+2,5*4+3*5+3,5*4+4*3+4,5*2+ +5)/25 = 75/25 = **3**

- $Var(Y_{medio})$ =[(1-3)^2+(1,5-3)^2*2+(2-3)^2*3+ +(2,5-3)^2*4+(3-3)^2*5+(3,5-3)^2*4+(4-3)^2*3+ +(4,5-3)^2*2+(5-3)^2)]/25 = 25/25 = **1**

Abbiamo così verificato "empiricamente" i valori, rispettivamente, della media e della varianza campionaria.

Dalla popolazione al campione. Per riassumere:

- **INFERENZA STATISTICA :** Insieme delle operazioni compiute dal ricercatore per "stimare" alcune caratteristiche (i **parametri**) di una popolazione, non interamente esplorabile, attraverso la selezione da questa di un sottoinsieme casuale di unità, detto **campione**.

- **PARAMETRO:** "Vero" valore (θ) assunto da una caratteristica misurata a livello di popolazione (somma, media, varianza, proporzione, coefficiente di regressione, coefficiente di correlazione, ecc.). Il parametro è, quasi sempre, incognito.

- **STIMATORE:** si dice **stimatore** qualunque **statistica** $T(X_1,X_2,...,X_n)$, ovvero una funzione applicata alle unità statistiche comprese nel campione, le cui determinazioni vengono utilizzate per ottenere una misura (stima puntuale) del parametro incognito θ . Pertanto, uno stimatore è esso stesso una variabile casuale e possiede una sua distribuzione, con i relativi momenti: valore atteso, varianza, ecc.

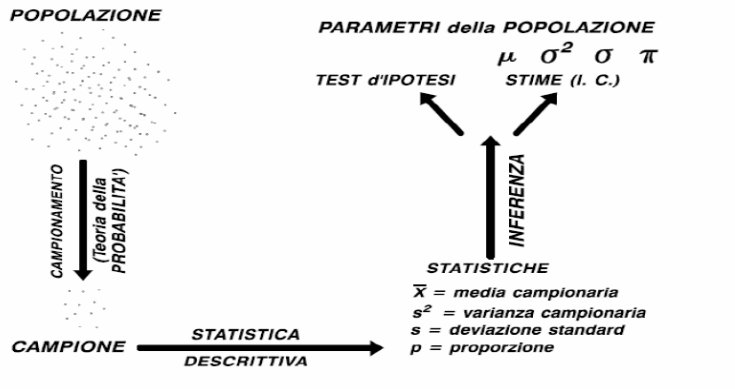
Errore di campionamento

- **ERRORE DI CAMPIONAMENTO**: Differenza tra il valore empirico dello stimatore (esempio: media) e il corrispondente valore che si sarebbe ottenuto analizzando la totalità delle unità statistiche della popolazione.
- L'errore di campionamento si verifica, come già accennato, perché quello osservato è solo un sottoinsieme (talvolta molto piccolo) delle unità della popolazione. L'errore di campionamento tende a diminuire all'aumentare della numerosità campionaria.

Stima puntuale, stima intervallare e test di ipotesi

- **OPERAZIONE DI STIMA:** l'insieme delle regole e attraverso le quali è stato ottenuto un particolare valore dello stimatore.
- La stima può riguardare soltanto il parametro, e in tal caso si parla di "stima puntuale", oppure un intervallo attorno al parametro stesso, e allora si parla di "stima intervallare".
- **TEST DI IPOTESI:** Processo decisionale, basato sul controllo di ipotesi statistiche effettuate sulla realtà osservata. Tale processo porta a rifiutare, oppure non rifiutare, una certa ipotesi (statistica) formulata sulla popolazione.

Schematicamente



Stima puntuale

- Sui dati del campione di cui disponiamo, il valore calcolato, empirico dello stimatore (di un certo parametro) è la **stima puntuale del parametro**.

- Solitamente si usa:
- La **media campionaria** per stimare la **media della popolazione**;
- La **varianza campionaria** per stimare la **varianza della popolazione**;
- La **differenza tra due medie campionarie** per stimare la **differenza tra due valori medi a livello di popolazione**;
- Eccetera

Stima per intervallo

- Con questa procedura di stima si determina un set di valori a partire dal campione che con una certa **probabilità,**
- **(1-α) %**, contiene il parametro incognito.
- **(1-α) % indica il livello (o grado) di confidenza**; l'intervallo che si ottiene è detto **intervallo di confidenza**.
- Gli estremi dell'intervallo dipendono dal campione estratto, quindi sono sottoposti a **variazioni casuali.**

Intervalli di confidenza

- Un intervallo di confidenza è quindi un **insieme di valori plausibili** per il parametro incognito sulla base dell'**evidenza empirica**.
- **Attenzione**: il livello di confidenza rappresenta il **grado di affidabilità della procedura**, **non** il grado di plausibilità **del risultato**, dovuto al singolo campione. La plausibilità del risultato è invece espressa dalla ampiezza dell'intervallo di confidenza.
- Generalmente, si usa come livello di confidenza il 95% (α=5%)

Ampiezza dell'intervallo di confidenza

- L'**ampiezza** dell'intervallo è molto importante:
- Quanto **più** l'intervallo è **ridotto**, tanto **maggiore** è il grado di **precisione della stima**.

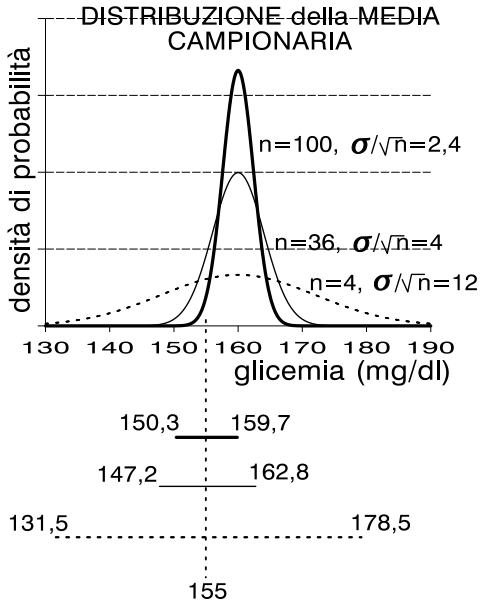
Ampiezza dell'intervallo di confidenza

- **L'ampiezza dell'intervallo** dipende quindi:
- dal grado di confidenza (1-α): **al diminuire di α** [cioè al **crescere del grado di confidenza** (1-α)], l'ampiezza dell'intervallo **aumenta**;
- dalla **variabilità** del fenomeno studiato: al **crescere della variabilità** dei dati che stiamo osservando cresce anche l'incertezza e quindi l'ampiezza dell'intervallo **aumenta**;
- dalla numerosità campionaria, *n*: al **crescere di *n*** aumenta la quantità di **informazione** disponibile e quindi l'ampiezza dell'intervallo **diminuisce**.

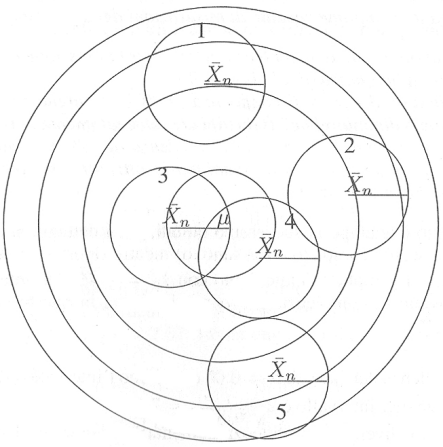
Come si può far diminuire l'ampiezza di un intervallo di confidenza?

- **L'ampiezza di un intervallo di confidenza diminuisce** se:

- **1**) diminuisce il **livello di confidenza** (1-α) (es. dal 99% al 95% al 90%)
- **2**) aumenta la **numerosità** del campione (es. da n=4 a n=36 a n=100)
- **3)** diminuisce la **variabilità nella popolazione** (es. da σ=100 a σ=36 a σ=4)



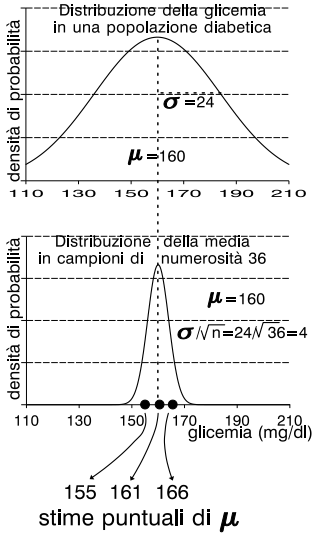
Esempio: “tiro al bersaglio” con la media campionaria



- Al centro del bersaglio c’è la media incognita μ. I cerchi con al centro i valori calcolati della media campionaria sono gli intervalli di confidenza di ampiezza costante, **il cui raggio dipende dalla numerosità campionaria, *n*, e dal livello di confidenza** α.
- Come si nota, alcuni intervalli (cerchi) non contengono il valore μ (si tratta degli intervalli 1, 2 e 5), mentre altri invece lo contengono (gli intervalli 3 e 4).
- Si può interpretare il livello di confidenza, (1-α), come la probabilità che i cerchi (intervalli) contengano il valore incognito μ.

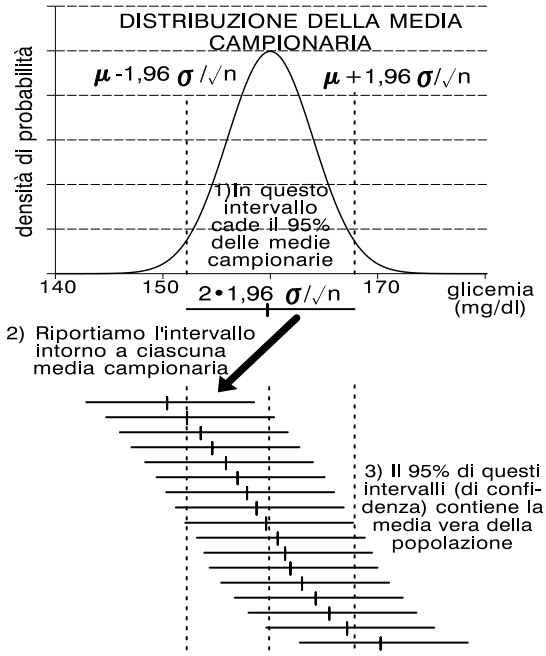
Stima puntuale e stima intervallare

- Una singola stima del valore medio è una **stima puntuale** (è un unico numero).
- Ma: il valore medio di un unico campione può essere una buona stima del valore medio di una popolazione, quando sappiamo che, prendendo diversi campioni della stessa popolazione, otterremo sempre un valore diverso?



Stima puntuale e stima intervallare

- Sappiamo che i campioni tratti da una stessa popolazione variano in modo sistematico.
- Oltre alla stima puntuale, possiamo conoscere anche una stima dell'intervallo in cui è "molto probabile" che cada il valore vero del parametro per la popolazione.
- Chiameremo questa valutazione è chiamata **stima intervallare.**



Stima puntuale e stima intervallare

- La **stima puntuale** fornisce un singolo valore. Da ciò consegue che:
- 1) questo valore non coincide quasi mai con il valore vero (parametro) della popolazione;
- 2) campioni diversi forniscono stime puntuali diverse.

- La **stima intervallare** fornisce un intervallo, che ha una predeterminata probabilità di contenere il valore vero del parametro. Pertanto, quest'intervallo ha una determinata probabilità prefissata (ad esempio, il 95%) di contenere il valore vero del parametro (della popolazione).

Esempio: intervallo di confidenza per la media, da una popolazione normale

- Intervallo di confidenza, con grado di confidenza $(1-\alpha)\%=$
- $= m + / - z_{\alpha/2} * [\sigma(X) / \sqrt{n}]$
- Dove:
- m = valore calcolato dello stimatore "media campionaria";
- $z_{\alpha/2}$ = valore critico della "normale standardizzata" in corrispondenza di α ;
- $[\sigma(X)/\sqrt{n}]$ = deviazione standard della v.c. "media campionaria (questa quantità è detta "errore standard" di X).

L'INTERVALLO DI CONFIDENZA, AL 95%, PER LA MEDIA CAMPIONARIA (da popolazione NORMALE)

- Se:
- $X \sim N(\mu,(\sigma^2/n))$ e quindi $Z \sim N(0,1)$ [normale standardizzata]
- Sappiamo che, con probabilità pari al 95%, l'intervallo compreso tra gli estremi:
- $[\, m - z_{\alpha/2} * [\sigma(X)/\sqrt{n}];\ m + z_{\alpha/2} * [\sigma(X)/\sqrt{n}] \,]$
- comprenderà il **vero valore del parametro** μ, non noto.

Un esempio

- Supponiamo di voler fare inferenza sul peso medio, alla nascita, di neonati di sesso maschile con età gestazionale di 39 settimane.
- **ESERCIZIO.**
- Sapendo che il peso alla nascita è una v.a. normale, con media incognita (μ) e deviazione standard (σ) nota, pari a 535 grammi, si calcoli l'intervallo al 95% per μ a partire da un campione casuale semplice, estratto dalla popolazione, di numerosità pari a 16.
- n = 16
- Media campionaria calcolata = 3434 g
- σ (nota) = 535 g

Un esempio

- Limite superiore dell' I.C. 95% = 3434 +1,96 *(535/√16) = 3172
- Limite inferiore dell' I.C. 95% = 3434 –1,96 *(535/√16) = 3696
- Intervallo di confidenza al 95%: [3172 ; 3696]
- **QUINDI:**
- Con probabilità pari al 95%, il peso medio alla nascita dei neonati maschi, nati alla 39ma settimana di gestazione, è un valore compreso tra **3172** e **3696**.

Intervalli di confidenza: un esempio tratto dalla letteratura

- Mostriamo un esempio tratto dalla ricerca medica, anche se gli intervalli di confidenza riportati sono stati ricavati con un'altra tecnica statistica;
- Dati tratti da un articolo apparso sul New England Journal of Medicine nel 1996: *Bone mineral density in women with depression*;
- L'ipotesi è che il soffrire o l'avere sofferto di depressione in passato provochi un **calo** della densità ossea nelle donne (meccanismi endocrini).

Intervalli di confidenza: un esempio tratto dalla letteratura

TABLE 3. BONE MINERAL DENSITY IN 24 DEPRESSED AND 24 NORMAL WOMEN.*

BONE MEASURED†	DEPRESSED WOMEN	NORMAL WOMEN	MEAN DIFFERENCE (95% CI)	P VALUE
Lumbar spine (anteroposterior) Density (g/cm²)	1.00±0.15	1.07±0.09	0.08 (0.02 to 0.14)	0.02
Lumbar spine (lateral)‡ Density (g/cm²)	0.74±0.09	0.79±0.07	0.05 (0.00 to 0.09)	0.03
Femoral neck Density (g/cm²)	0.76±0.11	0.88±0.11	0.11 (0.06 to 0.17)	<0.001
Ward's triangle Density (g/cm²)	0.70±0.14	0.81±0.13	0.11 (0.06 to 0.17)	<0.00
Trochanter Density (g/cm²)	0.66±0.11	0.74±0.08	0.08 (0.04 to 0.13)	<0.001
Radius Density (g/cm²)	0.68±0.04	0.70±0.04	0.01 (−0.01 to 0.04)	0.25

*Plus–minus values are means ±SD. CI denotes confidence interval.

79