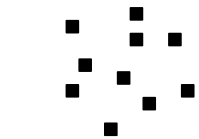
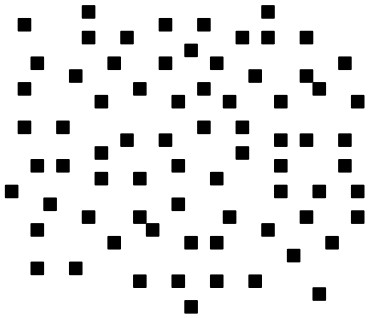
Intervallo di confidenza

SAPIENZA
UNIVERSITÀ DI ROMA

annarita.vestri@uniroma1.it



inferenza



Media

Stima puntuale di μ

Riportare sempre anche
la deviazione standard

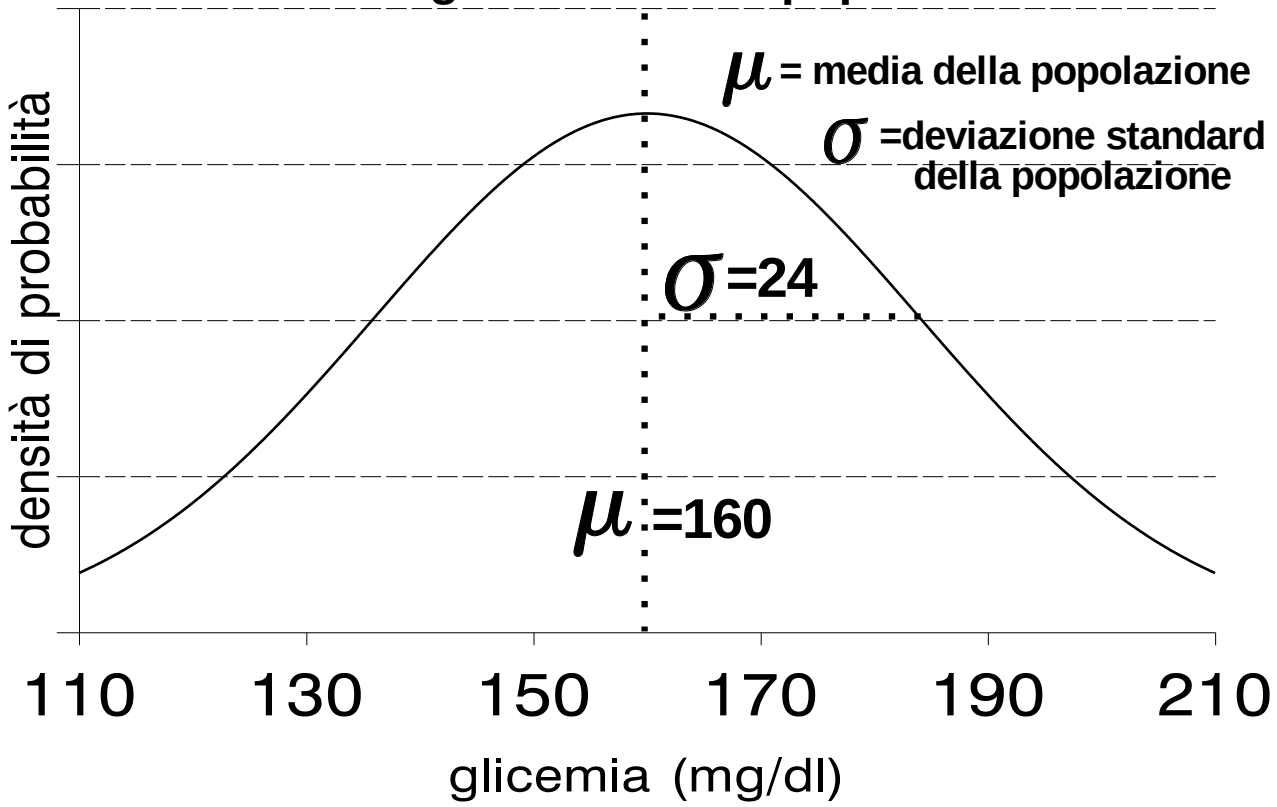
Media,
dev.standard,
numerosità

Qualche semplice
calcolo

Intervallo di confidenza
(stima intervallare di μ)

Su 20 intervalli di confidenza al 95%,
19 contengono μ, il valore vero della popolazione

Esempio di distribuzione normale:
distribuzione della glicemia in una popolazione diabetica



Dal momento che il campione viene estratto casualmente dalla popolazione, le conclusioni tratte da un campione possono essere errate.

L'inferenza statistica

1) cerca di stimare la probabilità di commettere errori

2) cerca di limitare la probabilità di commettere errori

INTERVALLO di CONFIDENZA

Lo scopo dell'inferenza statistica è la conoscenza dei **parametri** che caratterizzano una popolazione.

Per conoscere il parametro, però, dovremmo prendere in esame **tutte** le unità statistiche che costituiscono la popolazione; questo spesso è impossibile perché:

1. numerosità molto elevata
2. spesso la popolazione obiettivo è infinita

impossibile conoscere il **parametro**

Non potendo calcolare con esattezza il parametro, **ricorriamo ad una sua stima**.

La **statistica** (es $\bar{x}$, s) calcolata su un campione estratto dalla popolazione obiettivo è una **stima puntuale** del parametro della popolazione.

Questa stima puntuale del parametro non sarà mai identica al vero parametro della popolazione, ma sarà affetta da un **errore** per eccesso o per difetto.

La stima puntuale ha però un limite: non abbiamo modo di valutarne l'attendibilità.
Potrebbe infatti trattarsi di una stima molto vicina al valore ignoto, e quindi attendibile, oppure molto lontana da esso, risultando quindi del tutto fuorviante

Inoltre campioni diversi forniscono stime puntuali diverse, quindi l'aleatorietà che è alla base della costruzione del campione si ripercuote anche sui risultati dello stesso

In molte situazioni è preferibile **una stima intervallare** (cioè è preferibile indicare come stima del parametro un intervallo al posto di un *singolo punto* sull'asse dei valori) che esprima anche l'**errore associato alla stima** (precisione).

Intervallo di confidenza della media

Intervallo noto entro il quale, con una probabilità prefissata, si collochi la media ignota μ della popolazione

Intervallo di confidenza della media

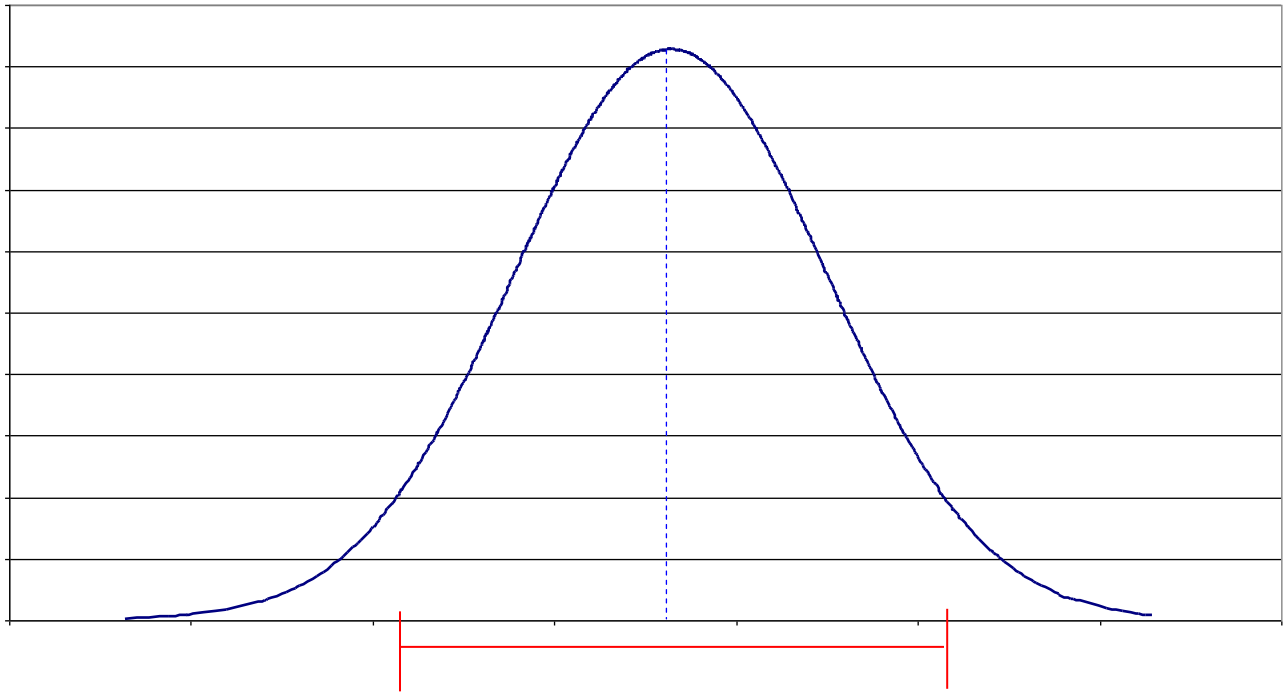
Partiamo dal concetto ovvio che noi vorremmo ottenere una stima il più possibile vicina al vero valore del parametro ignoto μ

Potrebbe comunque essere un compromesso soddisfacente l'individuazione un intervallo entro il quale le medie campionarie possano essere considerate "accettabili"

Intervallo di confidenza della media

Potremmo quindi decidere di *prendere per buone* una certa percentuale di medie campionarie comprese in un intervallo intorno al parametro μ, e *rifiutare* le restanti in quanto troppo lontane da μ

Intervallo di confidenza della media



Intervallo di confidenza della media

Che percentuale di medie vogliamo considerare accettabili?

Il 50 % ?

Il 75 % ?

Il 90 % ?

Intervallo di confidenza della media

Classicamente si considera accettabile una percentuale di medie campionarie compresa fra il 90 ed il 95 %

Ne consegue che ne viene rifiutata solo una percentuale compresa fra il 5 ed il 10 %

Intervallo di confidenza della media

Nella pratica, si fissa a priori la probabilità di avere una media “inaccettabile”, che viene indicata con α

Quindi α è normalmente compresa fra 0.05 e 0.10

Intervallo di confidenza della media

La probabilità α viene definita *probabilità di errore* ma anche area di rifiuto in relazione alla distribuzione di probabilità che si sta utilizzando (la Gaussiana o altro)

Intervallo di confidenza della media

Fissata infatti a priori la probabilità di errore α (ovvero di avere una media “inaccettabile”), è immediatamente definita la probabilità dell’evento complementare, ovvero che l'intervallo individuato comprenda una media campionaria accettabile, pari ad **1 - α**

Pertanto se $\alpha = 0.05$ sarà $1 - \alpha = 0.95$

Intervallo di confidenza della media

Rimarrà da stabilire l'ampiezza di tale intervallo (detto di confidenza), o meglio, il suo limite inferiore (L_i) ed il suo limite superiore (L_s)

Intervallo di confidenza della media

L’intervallo avrà una ampiezza simmetrica rispetto a μ per cui sarà dato, per ciascuna coda, dalla variabilità propria della distribuzione di medie campionarie (ES) moltiplicata per quel valore di Z (deviata standardizzata della curva Normale) relativo ad una probabilità pari ad $\alpha/2$

Intervallo di confidenza della media

Ricordando che

$$Z = \frac{\bar{x}_i - \mu}{ES}$$

Ci interessa trovare quel valore della Z che individua nella sua coda destra un'area (= probabilità) pari ad $\alpha/2$

Stessa cosa per la coda sinistra

Intervallo di confidenza della media

Stabilito quindi a priori il valore di α, usualmente pari a 0.05, si ricava immediatamente $\alpha/2 = 0.025$

Quindi, il corrispondente valore della deviata standardizzata $Z\alpha/2$, relativa a ciascuna delle due code della normale, si ricava dalle ben note tavole

Intervallo di confidenza della media

Sapendo che l'area a destra di Z deve essere pari a 0.025, ci interessa trovare il valore di $Z_{0.025}$

Quindi dalla tavola $Z_{0.025} = 1.96$

Intervallo di confidenza della media

Quindi, in formula…

$$pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

Intervallo di confidenza della media

Se $\alpha = 0.05$, ne consegue che $Z\alpha/2 = 1.96$

$$ES = \sigma(\bar{x}_i) = \frac{\sigma}{\sqrt{n}}$$

ES noto

$$\hat{ES} = \hat{\sigma}(\bar{x}_i) = \frac{s}{\sqrt{n}}$$

ES stimato

Intervallo di confidenza della media

Ma la formula trovata è solo quella che definisce l'intervallo di accettabilità di una ipotetica media campionaria (con una probabilità di errore $\alpha = 0.05$)

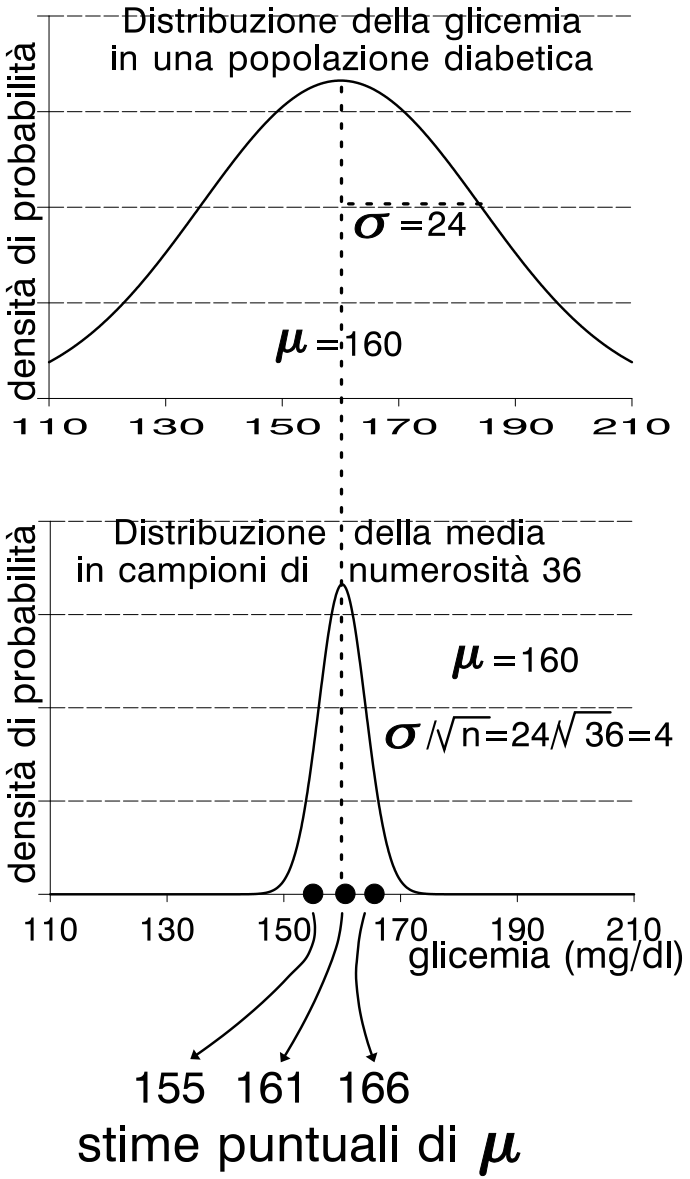
A noi interessa l'intervallo di confidenza (sempre con una probabilità di errore $\alpha = 0.05$) della media μ della popolazione

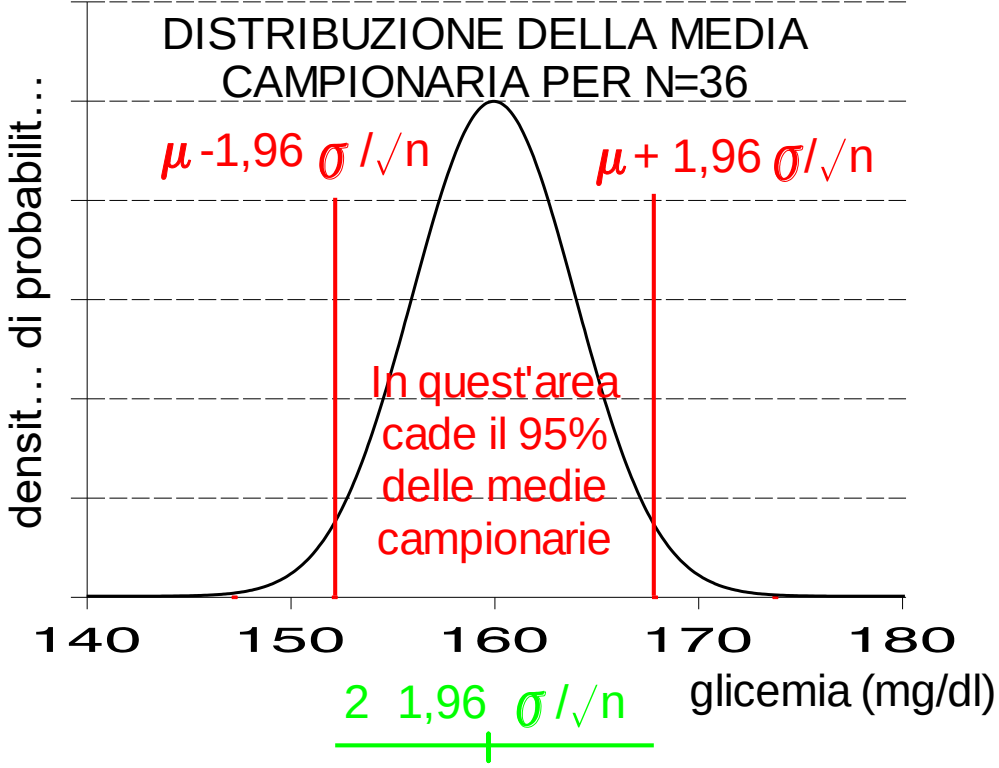
Intervallo di confidenza della media

Con passaggi matematici elementari, dalla formula precedente si arriva alla formula finale

$$pr\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

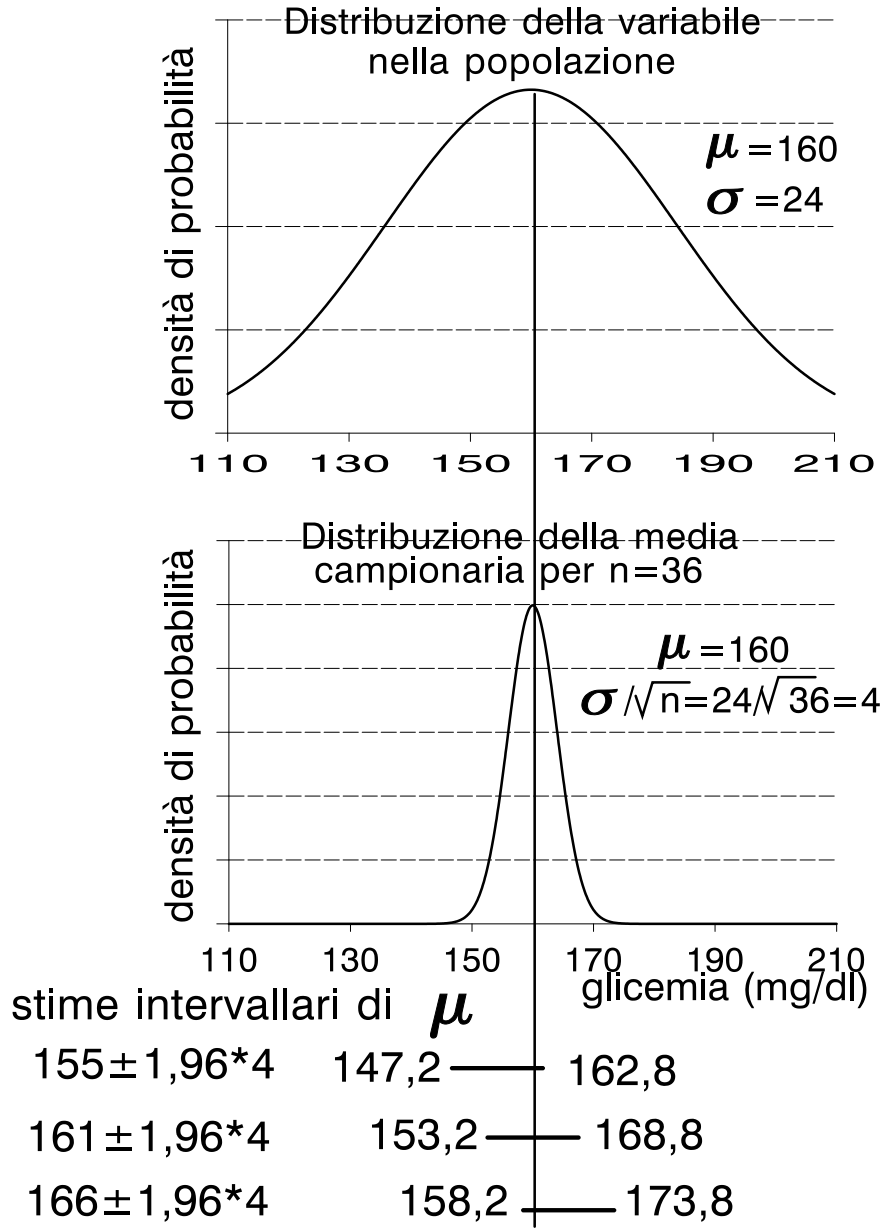
INTERVALLO DI CONFIDENZA


Distribuzione della glicemia
in una popolazione diabetica
densità di probabilità
σ =24
μ =160
110 130 150 170 190 210
Distribuzione della media
in campioni di numerosità 36
densità di probabilità
μ =160
σ /√n=24/√36=4
110 130 150 170 190 210
glicemia (mg/dl)
155 161 166
stime puntuali di μ

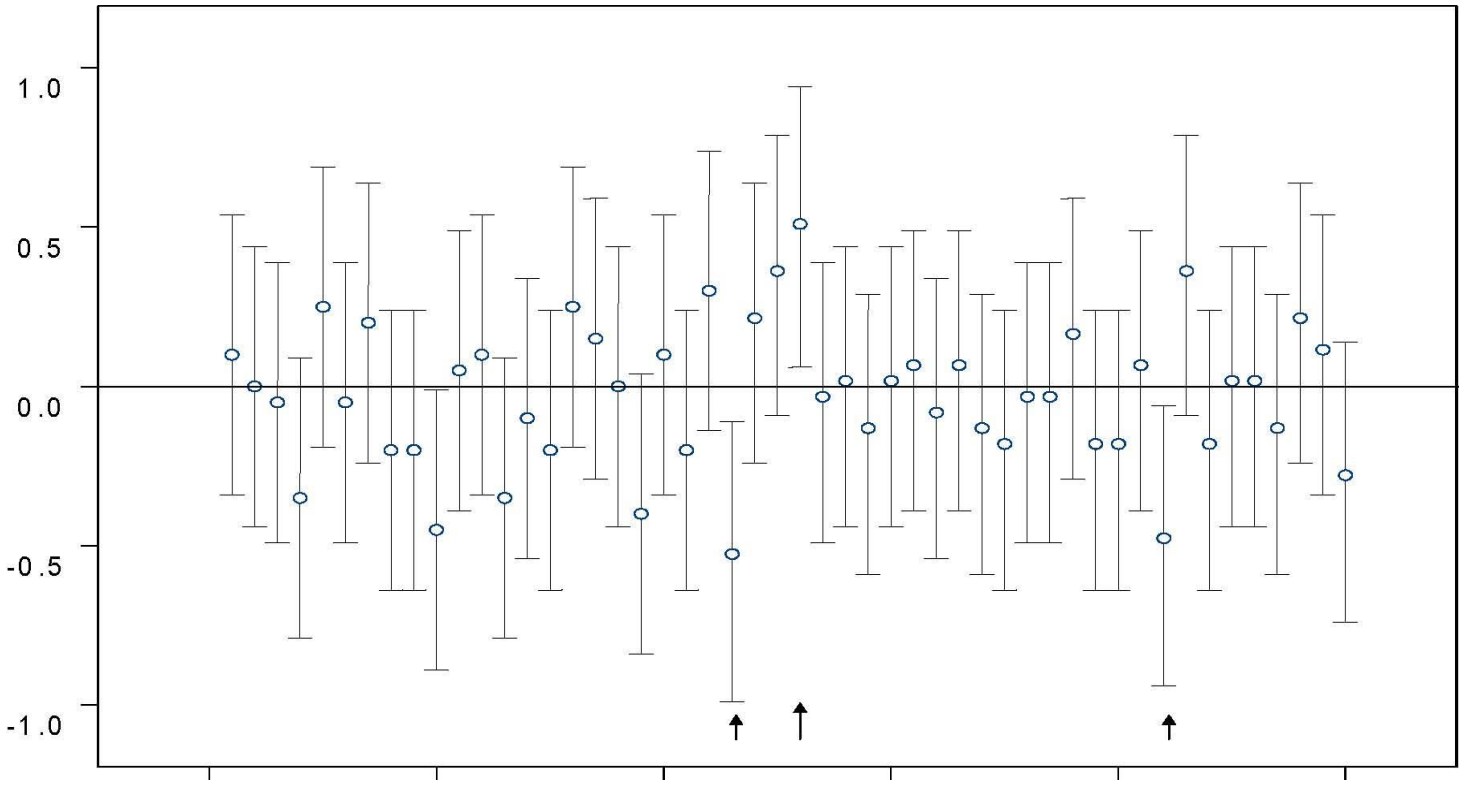

DISTRIBUZIONE DELLA MEDIA CAMPIONARIA PER N=36
μ -1,96 σ /√n
μ + 1,96 σ/√n
densit... di probabilit...
In quest'area cade il 95% delle medie campionarie
140
150
160
170
180
glicemia (mg/dl)


2 1,96 σ /√n


Se riportiamo quest'intervallo intorno a ciascuna media campionaria con la media esattamente al centro, il 95% di questi intervalli contiene la media vera della popolazione


Distribuzione della variabile nella popolazione
densità di probabilità
μ =160
σ =24
110 130 150 170 190 210
Distribuzione della media campionaria per n=36
densità di probabilità
μ =160
σ /√n=24/√36=4
110 130 150 170 190 210
glicemia (mg/dl)
stime intervallari di μ
155±1,96*4 147,2 162,8
161±1,96*4 153,2 168,8
166±1,96*4 158,2 173,8

Estrazione di 50 campioni di numerosità 20 da una distribuzione gaussiana con μ=0 e δ=1. Le barre rappresentano gli intervalli di confidenza al 95% per tutte le 50 medie campionarie calcolate.
Dati i 50 campioni dell'esempio seguente, osserviamo che soltanto in tre casi (6% dei campioni) l'intervallo di confidenza non comprende la vera media di popolazione.



La **stima puntuale** fornisce un singolo valore. Tuttavia:

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 della popolazione. Pertanto:

1) quest'intervallo ha una determinata probabilità (in genere, il 95%) di contenere il valore vero (parametro) della popolazione;

2) gli intervalli ottenuti da campioni diversi in genere si sovrappongono.

Per intervallo di confidenza di un parametro Θ della popolazione, intendiamo un intervallo delimitato da due limiti L_{inf} (limite inferiore) ed L_{sup} (limite superiore) che abbia una definita probabilità $(1-\alpha)$ di contenere il vero parametro della popolazione:

$$p(L_{inf} < \Theta < L_{sup}) = 1-\alpha$$

dove:

$1-\alpha$ = grado di confidenza

α = probabilità di errore

$$pr\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

invertendo le due disuguaglianze interne alla parentesi:

$$pr\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

INTERVALLO DI CONFIDENZA

$$pr\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

L_s

(LIMITE SUPERIORE DELL'INTERVALLO)

L_i

(LIMITE INFERIORE DELL'INTERVALLO)

L’ampiezza dell’intervallo dipende:

- **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.

L'intervallo di confidenza diminuisce se

1) diminuisce il **livello di confidenza** $(1-\alpha)$

(dal 99% al 95% al 90%)

2) aumenta la **numerosità** del campione

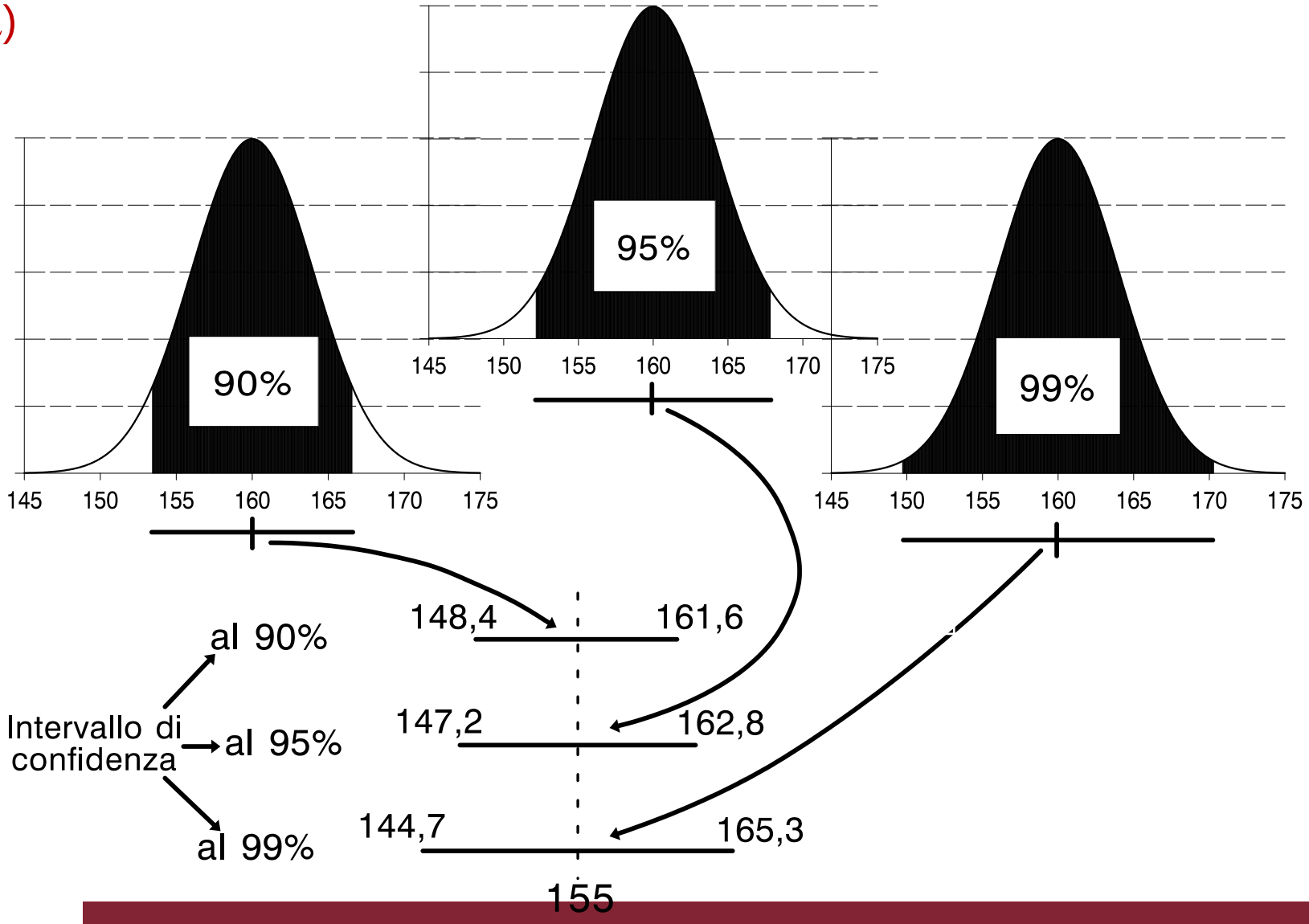
(da n=4 a n=36 a n=100)

3) diminuisce la **variabilità nella popolazione**

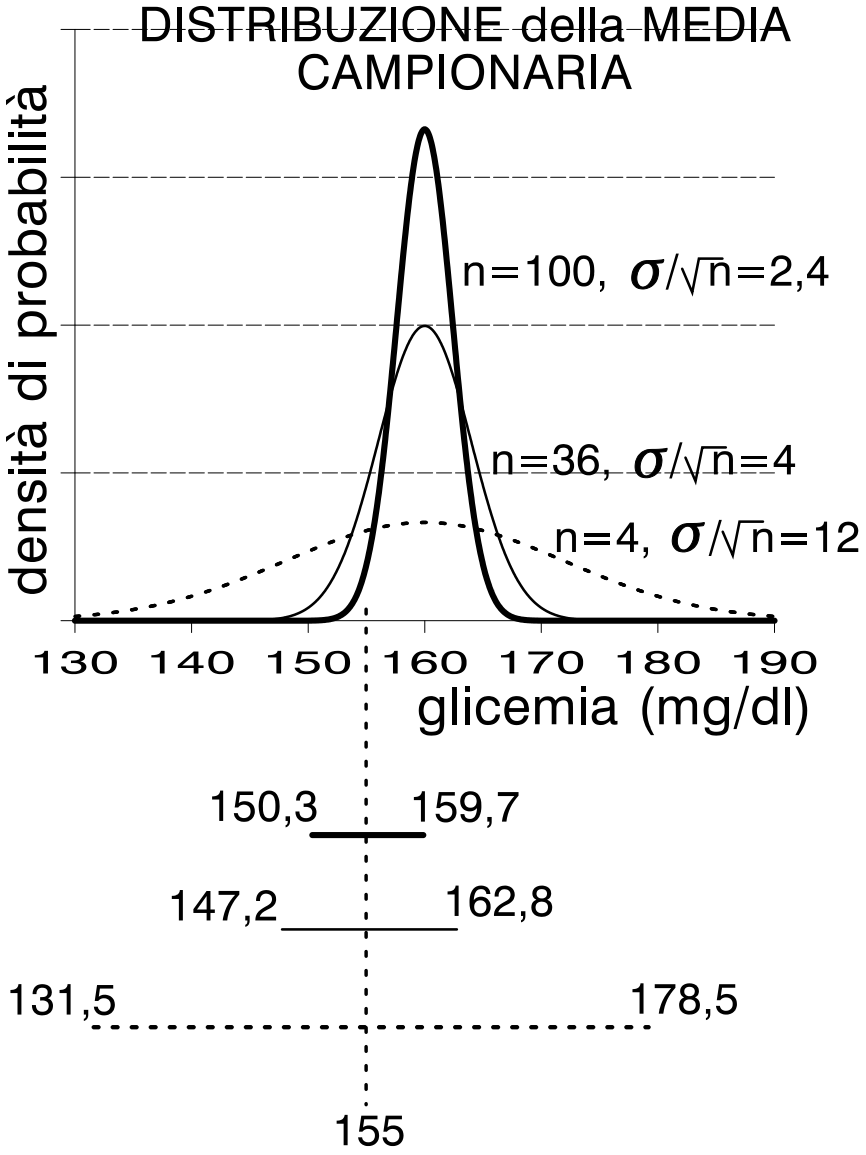
(da σ=100 a σ=36 a σ=4)

L'intervallo di confidenza diminuisce se diminuisce il **livello di confidenza**

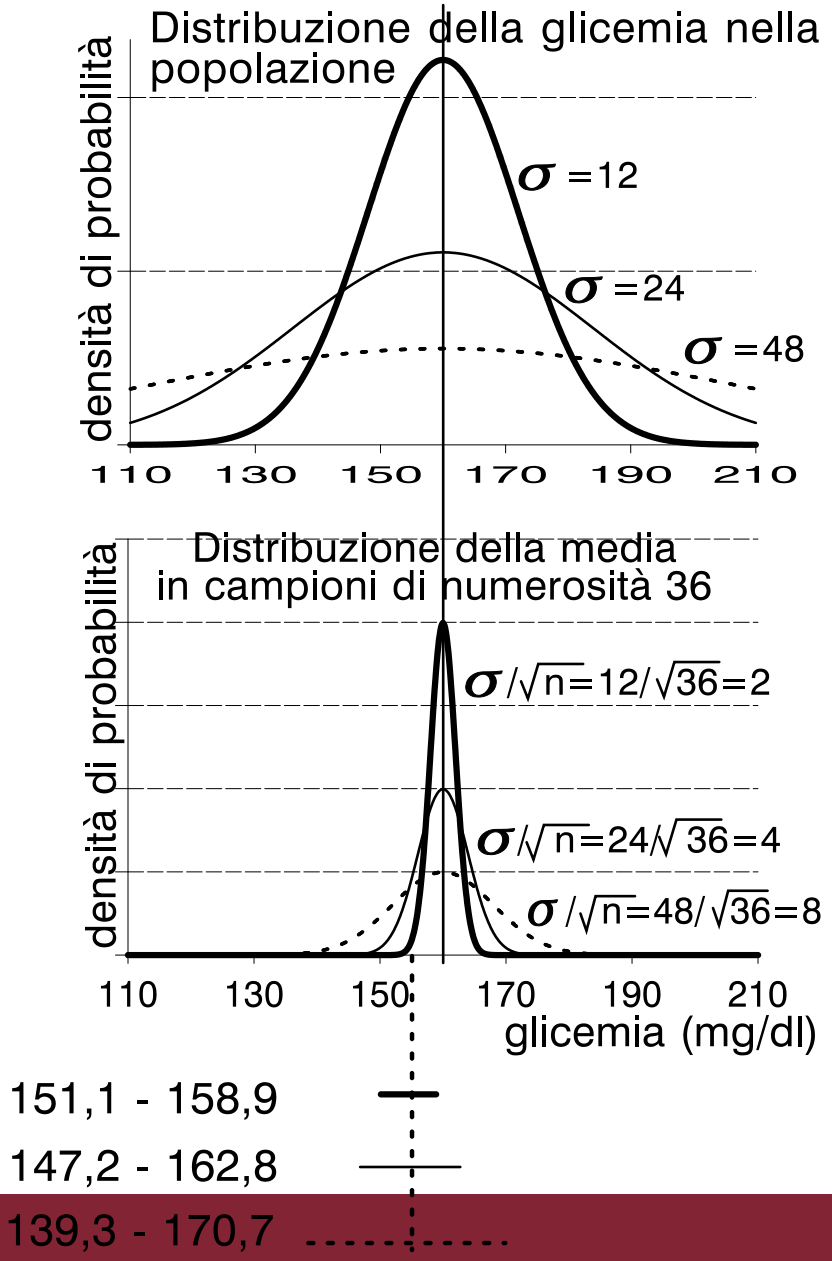
(1-α)



IC diminuisce se **aumenta** la **numerosità** del campione



IC diminuisce se diminuisce la **variabilità nella popolazione**



$$\bar{x} \pm z_{\alpha/2} \cdot ES(\bar{x})$$

la **probabilità d'errore α** determina il valore del coefficiente z:

1-α	α/2	$z_{\alpha/2}$
0.90	0.05	1.64
0.95	0.025	1.96
0.99	0.005	2.58

RIASSUMENDO…

La **stima puntuale** fornisce un singolo valore. Tuttavia:

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:

1. quest'intervallo ha una determinata probabilità (in genere, il 95%) di contenere il valore vero (parametro) della popolazione;
2. Il metodo generale per la costruzione dell'intervallo di confidenza al (1-α) è:

$$\bar{x} \pm z_{\alpha/2} \cdot ES(\bar{x})$$

Deviata gaussiana Standard: Aree per z>+z* (o per z<-z*)

z*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831

Deviata gaussiana Standard: Aree Per z>+z* (o per z<-z*)

z*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2-4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
3.0	.00135	.00097	.00069	.00048	.00034	.00023	.00016	.00011	.00007	.00005
4.0	.00003	.00002	.00001	.00001	.00001	.00000	.00000	.00000	.00000	.00000

Esempio: Calcolo dell'intervallo di confidenza della media di una popolazione

Problema: Qual è l'intervallo di confidenza al 95% della media del peso di una popolazione, se la media di un campione di 16 soggetti è pari a 75 Kg? Nella popolazione il peso è distribuito normalmente con deviazione standard pari a 12 Kg.

Dati: $x = 75$ Kg $\quad \sigma = 12$ Kg $\quad n = 16$ $\quad 1-\alpha = 95\%$ $\quad z_{\alpha/2} = 1{,}96$

Formula da utilizzare: $I.C._{95\%} = x \pm z_{\alpha/2} \cdot \sigma/\sqrt{n} = x \pm z_{\alpha/2} \cdot E.S.$

I passo: calcolo l'errore standard

$E.S. = \sigma/\sqrt{n} = 12/\sqrt{16} = 12/4 = 3$ Kg

II passo: calcolo l'intervallo di confidenza

$I.C._{95\%} = x \pm z_{\alpha/2} \cdot E.S. = 75 \pm 1{,}96 \cdot 3 =$

80,88 Kg

69,12 Kg

L'intervallo che va da 69,12 Kg (limite inferiore) a 80,88 Kg (limite superiore) ha 95 probabilità su 100 di contenere la media vera della popolazione.

E se non conosco σ, la deviazione standard della popolazione?

Posso usare s (dev. standard del campione) come stima di σ

Se la numerosità campiona-ria è sufficientemente grande (n≥30), s è una stima precisa di σ.

$$I.C. = x \pm Z_{\alpha/2} * s / \sqrt{n}$$

Se la numerosità campionaria è piccola (n<30)

1.non può essere applicato il teorema del limite centrale

2. s non è una buona stima di σ

?

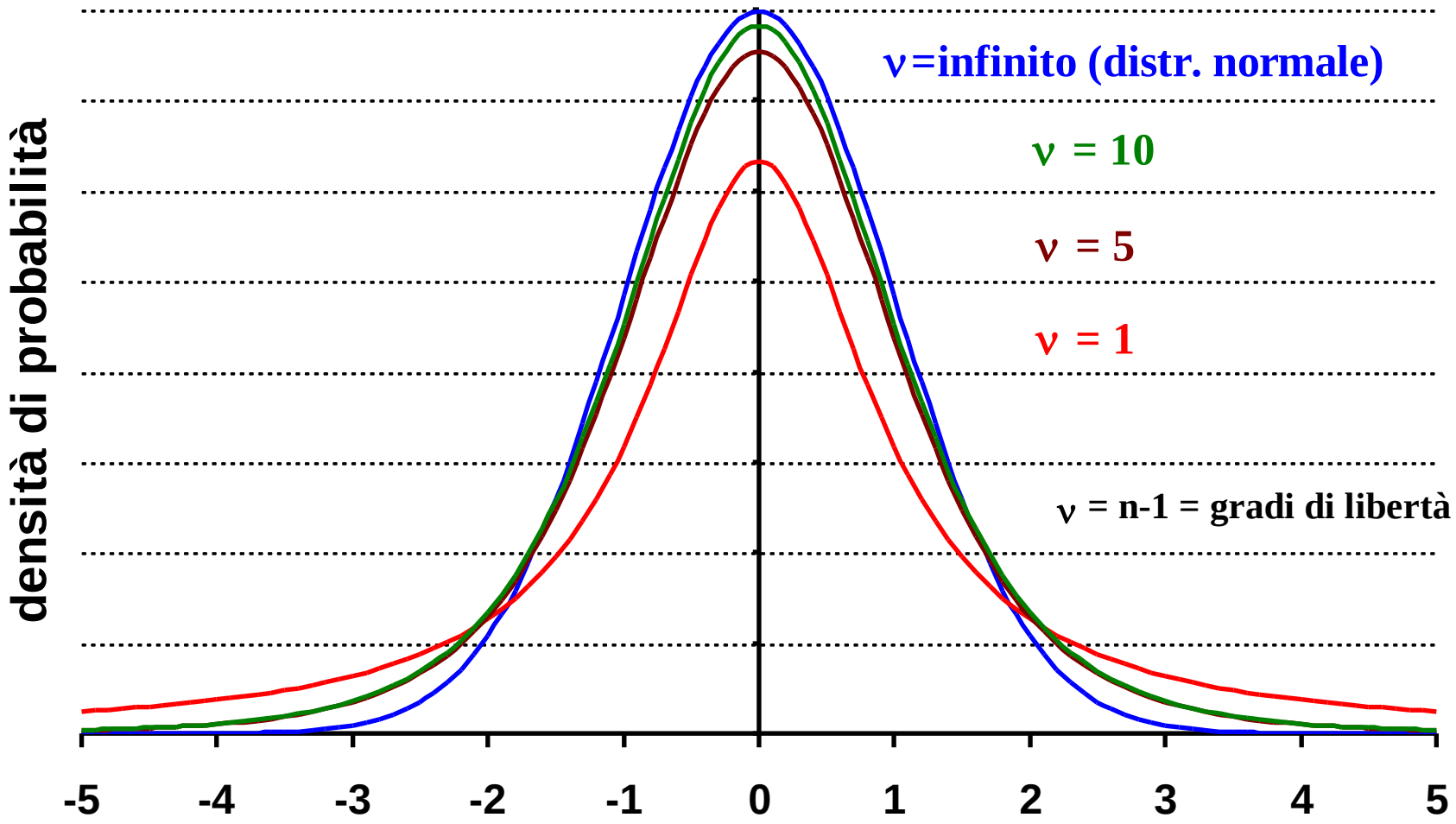
Per poter fare inferenze sulla media nel caso di piccoli campioni è necessario **assumere** che la variabile in studio abbia una **distribuzione approssimativamente normale**.

Sotto tale assunzione si può utillizzare la **distribuzione t di Student** con ν=(n-1) gradi di libertà

Nel caso di piccoli campioni l'intervallo di confidenza della media diventa quindi:

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Distribuzione T di Student



$$z = \frac{x-\mu}{\sigma/\sqrt{n}} \qquad t = \frac{x-\mu}{s/\sqrt{n}}$$

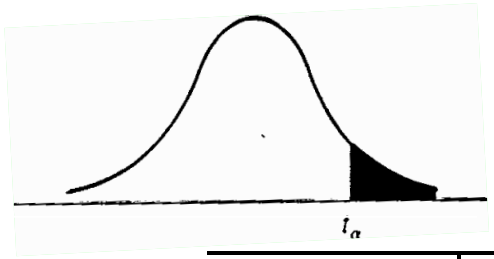


Tavola dei valori della funzione t di Student in funzione dei gradi di libertà e della probabilità in una coda della distribuzione (.100, .050, .025, .010, .005)

DEGREES OF FREEDOM	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

Esempio: Calcolo dell'intervallo di confidenza della media di una popolazione

Problema: Qual è l'intervallo di confidenza al 95% della media del peso di una popolazione, se in un campione di 100 soggetti la media (x) è pari a 75 Kg e la deviazione standard (s) è pari a 12 Kg.

Dati: x = 75 Kg s = 12 Kg n = 100

Il livello di confidenza è pari al 95%. Pertanto la probabilità di errore (α = alfa) è 100% - 95% = 5%.

Formula da utilizzare:

Poiché la numerosità campionaria è maggiore di 30, posso usare la distribuzione z.

Essendo il livello di confidenza del 95%, z=1,96

Pertanto posso utilizzare la seguente formula: $I.C._{95\%} = x \pm 1{,}96 \cdot s/\sqrt{n}$

I passo: calcolo l'errore standard (= la deviazione standard della media campionaria)

E.S. (Errore Standard) = $s/\sqrt{n} = 12/\sqrt{100} = 12/10 = 1{,}2$ Kg

II passo: calcolo l'intervallo di confidenza

$I.C._{95\%} = x \pm 1{,}96 \cdot E.S. = 75 \pm 1{,}96 \cdot * 1{,}2 =$

$75 + 1{,}96*1{,}2 = 77{,}35$ Kg

$75 - 1{,}96*1{,}2 = 72{,}65$ Kg

L'intervallo che va da 72,65 Kg (limite inferiore) a 77,35 Kg (limite superiore) ha 95 probabilità su 100 di contenere la media vera della popolazione.

Intervallo di confidenza della media

Esempio

Nel Reparto Nipiologico di un grande Ospedale romano (San XYZ) da alcuni anni si osserva una elevata incidenza di neonati immaturi (peso alla nascita < 2500 gr), con conseguente difficoltà di assicurare a tutti un adeguato ricovero in incubatrice

Intervallo di confidenza della media

L’Ospedale in oggetto ha per bacino di utenza un’area molto popolare ed accoglie, quindi, una ampia quota di extracomunitari, soprattutto di etnie asiatiche (cinesi) e nordafricane (neri)

Intervallo di confidenza della media

Poiché è noto che tali etnie presentano un peso medio alla nascita sensibilmente inferiore a quello della razza bianca, anche per neonati a termine, si ipotizza che il picco di immaturi osservato sia dovuto alla massiccia presenza di individui di origine extracomunitaria

Intervallo di confidenza della media

Tali neonati in realtà non avrebbero necessità del ricovero in incubatrici, ma i medici avendo come riferimento clinico il cut-off riferito alla razza bianca (2500 gr), sono indotti a farli rientrare nella categoria degli immaturi ed a trattarli come tali

Intervallo di confidenza della media

Ci si può chiedere:

qual è il peso medio dei neonati di etnia cinese?

Intervallo di confidenza della media

Per rispondere a tale domanda si estrae un campione casuale semplice, di numerosità n = 64, relativo ai neonati di etnia cinese venuti alla luce nel biennio 2015-16 nell'ospedale in oggetto

Intervallo di confidenza della media

I risultati ottenuti dal campione sono i seguenti

$$\bar{x} = 2750\,gr$$

$$s = 1600\,gr$$

$$n = 64$$

Intervallo di confidenza della media

Sulla base di questi dati calcoliamo l'errore standard

$$\hat{E}S = \frac{s}{\sqrt{n}} = \frac{1600}{\sqrt{64}} = \frac{1600}{8} = 200$$

Intervallo di confidenza della media

Pertanto l'intervallo di confidenza sarà

$$\bar{x} \pm z_{\alpha/2} \cdot ES$$

$$2750 \pm 1.96 \cdot 200$$

$$2358 \leq \mu \leq 3142$$

Intervallo di confidenza della media

Quindi si può affermare che il peso alla nascita di neonati di etnia cinese venuti alla luce nell'Ospedale S. XYZ nel biennio 2015-16 è compreso fra 2358 e 3142

Tale affermazione ha una probabilità di essere corretta pari a $1-\alpha = 0.95$

In altre parole, se effettuassimo 20 campioni simili al precedente, solo 1 su 20 (= 0.05) non comprenderebbe il parametro ignoto

esempio

Congenital malformation was seen in 2037 (4.97%) of exposed offspring and in 9443 (4.78%) of unexposed offspring.
The adjusted risk for congenital malformation was 4.98% in exposed offspring and 4.96% in unexposed offspring (risk difference 0.02% (95% confidence interval –0.26% to 0.30%).
The corresponding risk differences were 0.16% (–0.23% to 0.56%) with vaccination during the first trimester and 0.10% (–0.41% to 0.62%) with vaccination in the first eight weeks.