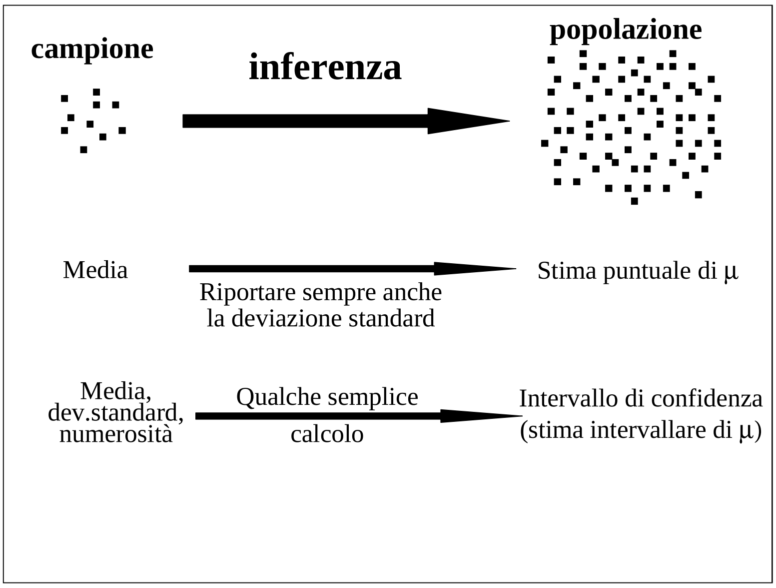
Intervallo di confidenza

Sezione di Epidemiologia e Statistica Medica,
Università di Verona

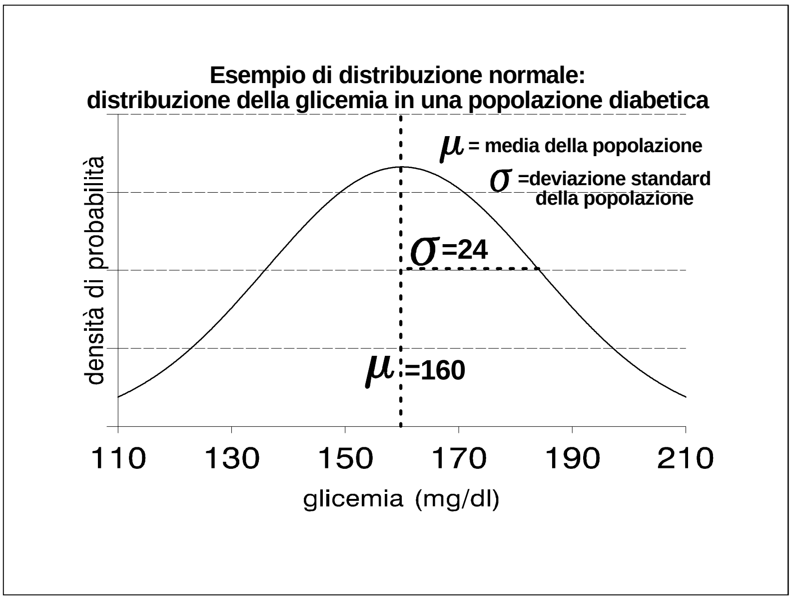


Dal momento che il campione viene estratto casualmente dalla popolazione, le conclusioni tratte da un campione possono essere errate.

L'inferenza statistica viene fatta "con umiltà":

1) si cerca di stimare la probabilità di commettere errori

2) si 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

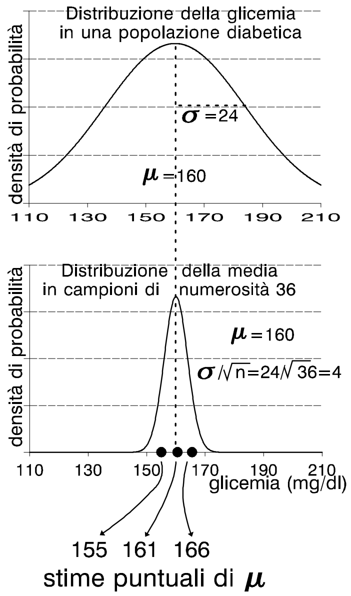
impossibile conoscere il **parametro**

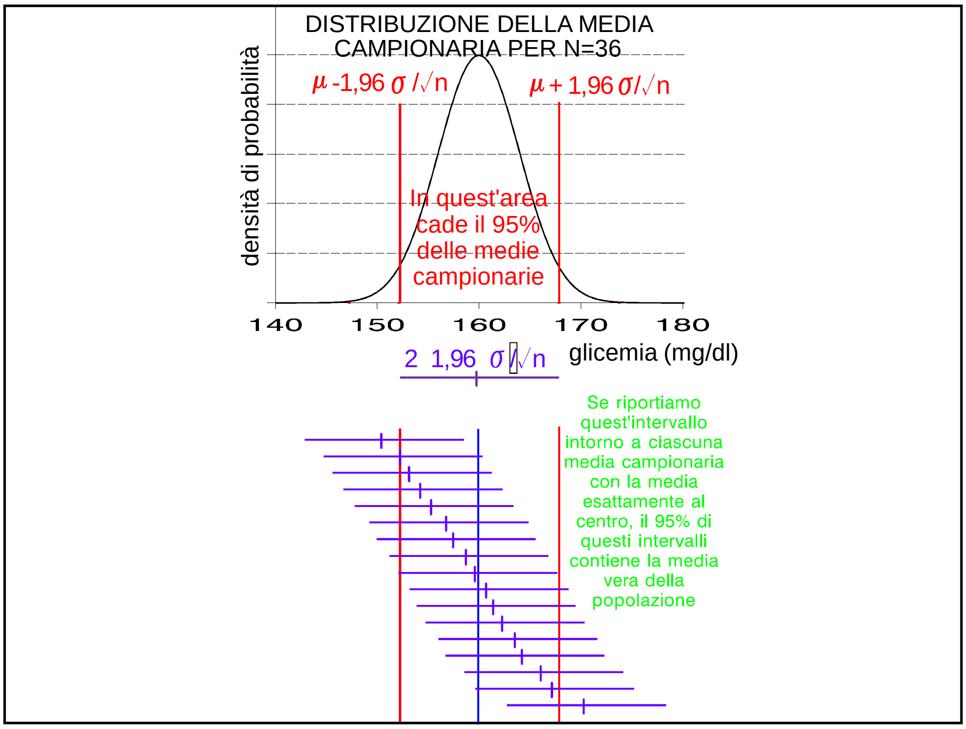
Non potendo calcolare con esattezza il parametro, **ricorriamo ad una sua stima**.

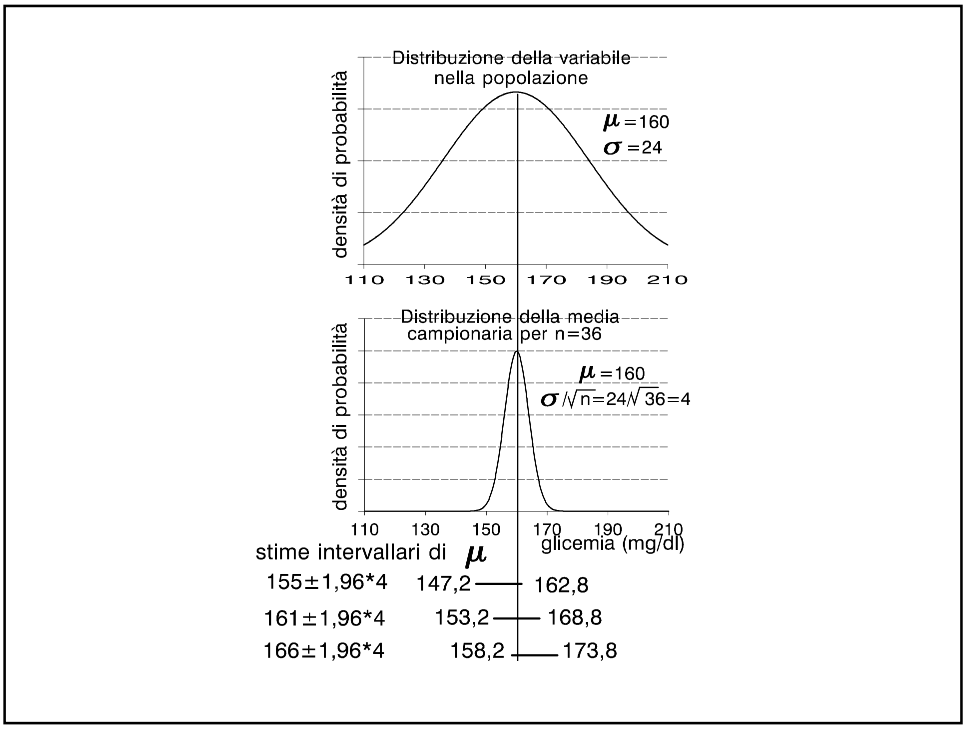
La **statistica** (es $\overline{X}$, s) calcolata su un campione estratto dalla popolazione obiettivo è una **stima puntuale** del parametro della popolazione.

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.

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).




DISTRIBUZIONE DELLA MEDIA
CAMPIONARIA PER N=36
densità di probabilità
μ -1,96 σ /√n
μ + 1,96 σ/√n
In quest'area
cade il 95%
delle medie
campionarie
140 150 160 170 180
2 1,96 σ /√n
glicemia (mg/dl)
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
μ =160
σ /√n=24/√36=4
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

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.

INTERVALLO di CONFIDENZA: DEFINIZIONE

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

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

e, riarrangiando le due disuguaglianze interne alla parentesi:

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

INTERVALLO DI CONFIDENZA

SESM

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

L_s

(LIMITE SUPERIORE DELL'INTERVALLO)

L_i

(LIMITE INFERIORE DELL'INTERVALLO)

SESM

L'intervallo di confidenza diminuisce se

1) diminuisce il **livello di confidenza** (1-α)

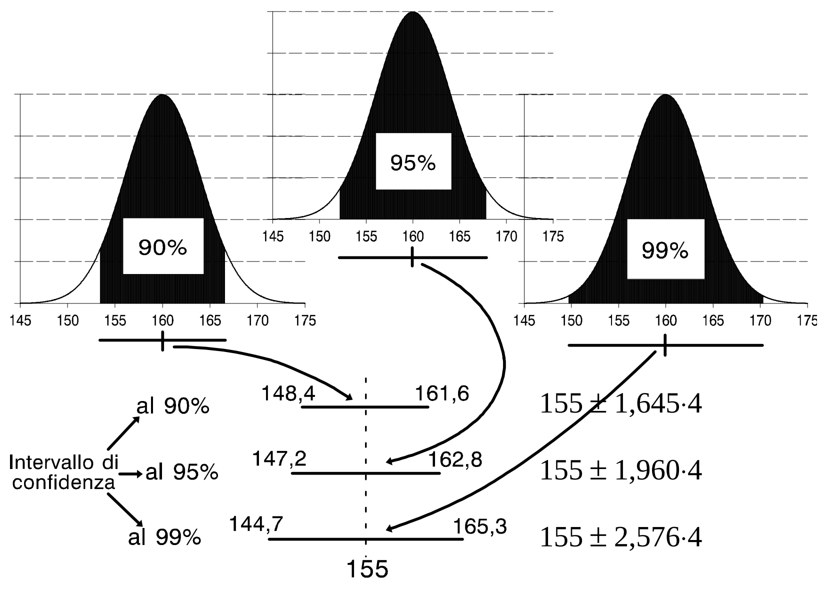
   (dal 99% al 95% al 90%)

2) aumenta la **numerosità** del campione

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

3) diminuisce la **variabilità nella popolazione**

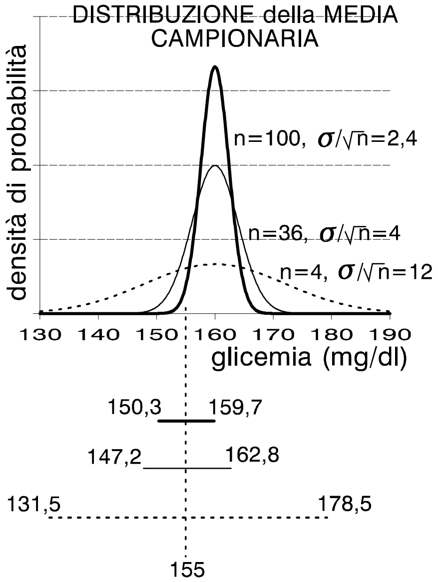
   (da σ=100 a σ=36 a σ=4)

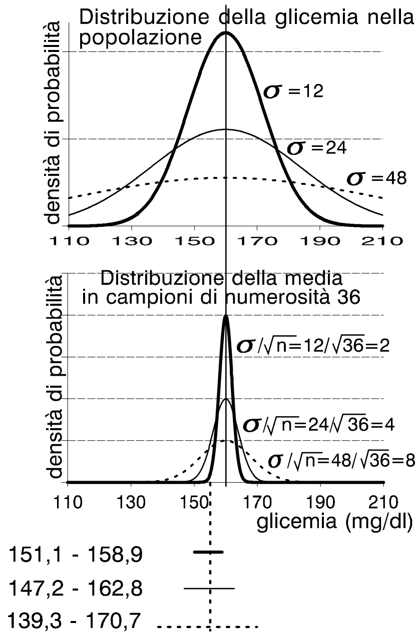


$$\bar{x} \pm z_{\alpha/2} \cdot ES(\bar{x})$$

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

$1-\alpha$	$\alpha/2$	$z_{\alpha/2}$
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:

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

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: $\bar{x}$ = 75 Kg σ = 12 Kg   n = 16   1-α= 95% $z_{\alpha/2}$ = 1,96

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

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

$I.C._{95\%} = \bar{x} \pm z_{\alpha/2} \cdot E.S. = 75 \pm 1{,}96 \cdot 3 =$ 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 campionaria è sufficientemente grande (n≥30), s è una stima precisa di σ.

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

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

?

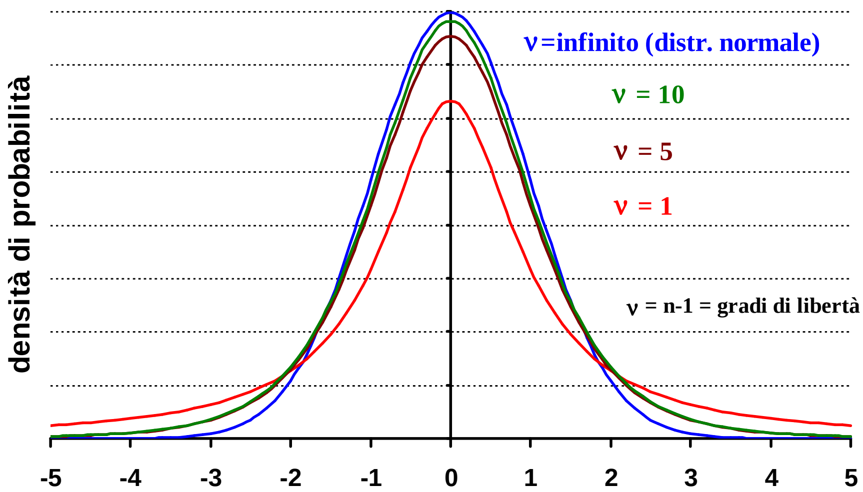
Per poter fare inferenze sulla media nel caso di piccoli campioni è necessario **assumere** che la variabile in studio abbia una **distribuzione approssimativamente normale**.

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



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

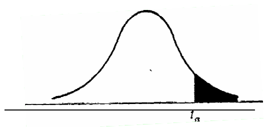


Tavola dei valori della funzione t di Student in funzione dei gradi di libertà e della probabilità in una coda della distribuzione (.100, .050, .025, .010, .005)

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

Riassumendo:

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

σ nota $\Rightarrow$ $\bar{x} \pm Z_{\alpha/2} * \sigma / \sqrt{n}$

σ ignota **n ≥ 30** $\Rightarrow$ $\bar{x} \pm Z_{\alpha/2} * s / \sqrt{n}$

σ ignota **n < 30** $\Rightarrow$ $\bar{x} \pm t_{\alpha/2,\nu} * s / \sqrt{n}$

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.

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 e la deviazione standard è pari a 12 Kg?

Dati: x = 75 Kg   s = 12 Kg   n = 16 $1-\alpha = 95\%$ $t_{15, \alpha/2} = 2{,}131$

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

I passo: calcolo l'errore standard
E.S. = $s / \sqrt{n} = 12/\sqrt{16} = 12/4 = 3$ Kg

II passo: calcolo l'intervallo di confidenza

$I.C._{95\%} = x \pm t_{15, \alpha/2} \cdot E.S. = 75 \pm 2{,}131 \cdot 3 =$
81,39 Kg
68,61 Kg

L'intervallo che va da 68,61 Kg (limite inferiore) a 81,39 Kg (limite superiore) ha 95 probabilità su 100 di contenere la media vera della popolazione.