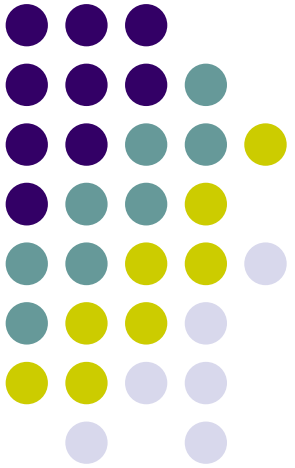
Elementi di Inferenza Statistica

Stima puntuale ed intervallare

V Scuola Estiva AISV

La statistica come strumento di analisi nelle scienze umanistiche e comportamentali

Soriano nel Cimino (VT), 6 Ottobre 2009



Pier Francesco Perri
Dipartimento di Economia e Statistica - UNICAL
pierfrancesco.perri@unical.it

Campione e Inferenza

Conoscere alcune caratteristiche incognite della popolazione oggetto di studio è un'esigenza che accumuna tutte le scienze empiriche.

La conoscenza **"esatta"** della popolazione si realizza solo quando è possibile rilevare il fenomeno su tutte le unità elementari che la compongono.

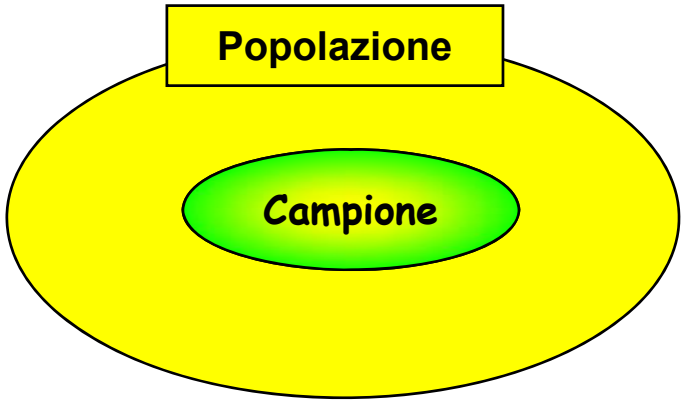
- **L'indagine censuaria** presenta alcune difficoltà operative legate soprattutto ai fattori

tempo

costi

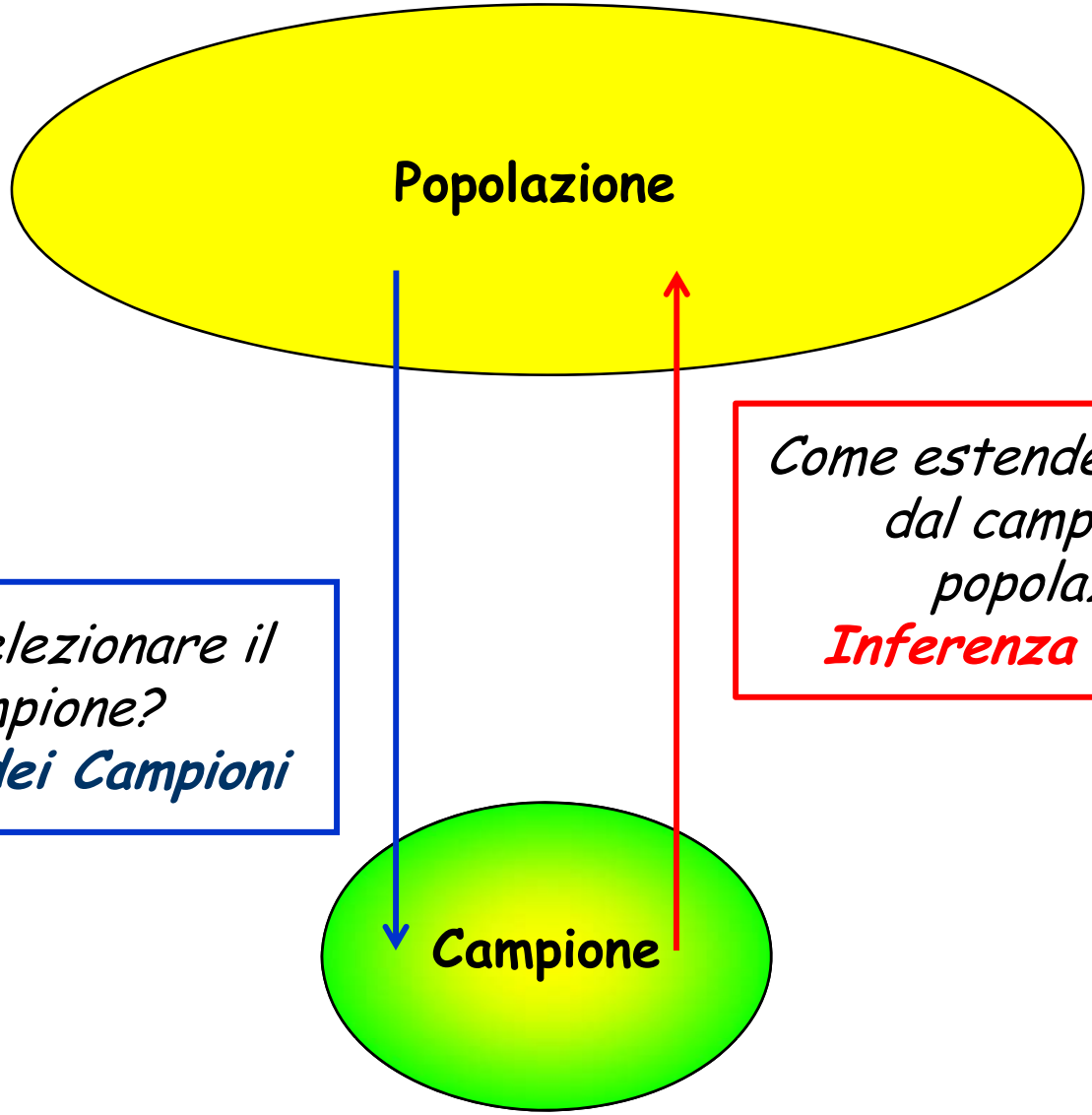
distruzioni delle unità

- **L'indagine campionaria** analizza solo un sottoinsieme di unità della popolazione chiamato **campione**.



Trattandosi di un'analisi parziale della realtà si perviene ad una **stima,** più o meno esatta, delle caratteristiche della popolazione che si intendono studiare.

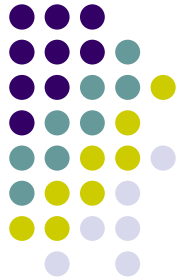
- ✓ come selezionare il campione?
- ✓ come estendere i risultati dal campione alla popolazione?



Come selezionare il campione?
Teoria dei Campioni

Come estendere i risultati dal campione alla popolazione?
Inferenza Statistica

Inferenza ed errori



Il meccanismo inferenziale attraverso il quale si risale dal particolare (il campione) al generale (popolazione) è un

"processo d'azzardo"

nel senso che non è possibile fare generalizzazioni assolutamente certe.

Le decisioni e i risultati che scaturiscono da tale processo comportano l'assunzione di un rischio dovuto sia alla limitatezza delle informazioni, sia alla natura casuale del campione:

- possono essere diversi a seconda del campione selezionato
- risultano maggiormente attendibili quando la dimensione del campione è elevata

Inferenza ed errori

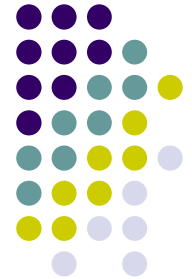
Tuttavia, se le procedure inferenziali utilizzate hanno una solida base metodologica, il grado di incertezza legato ai risultati può essere controllato e misurato in termini di probabilità.

Pertanto, l'Inferenza Statistica fornisce, non solo i metodi per risalire dal campione alla popolazione, ma anche per misurare il grado di incertezza insito nel procedimento.

Fra gli errori ci sono quelli che puzzano di fogna, e quelli che odorano di bucato.

Cesare Pavese (1908-1950)

L' Inferenza Statistica



L'inferenza statistica è strutturata in tre grandi branche:

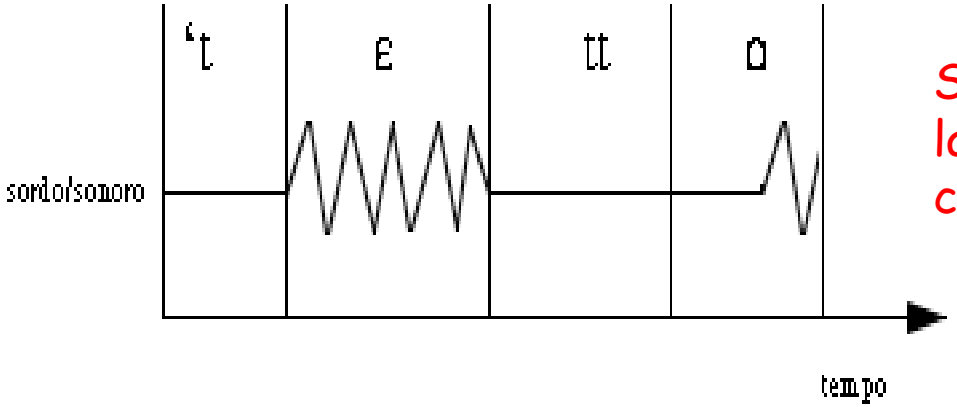
- **Stima puntuale:** a partire dalle osservazioni campionarie sul fenomeno oggetto di studio si determina un valore della caratteristica incognita (*parametro*) della popolazione
- **Stima intervallare:** a partire dalle osservazioni campionarie si determina un intervallo contenente il parametro incognito della popolazione
- **Verifica di ipotesi:** sulla base dei dati campionari si decide se un'ipotesi su un parametro della popolazione è vera o falsa

La stima puntuale

In Fonetica è noto che alcuni suoni si caratterizzano per un breve periodo di sordità che si presenta durante ed anche dopo l'esplosione di alcune consonanti soprattutto occlusive.

Ciò può causare un ritardo nell'attacco della sonorità del segmento seguente che viene chiamato **VOT** (Voice Onset Time).

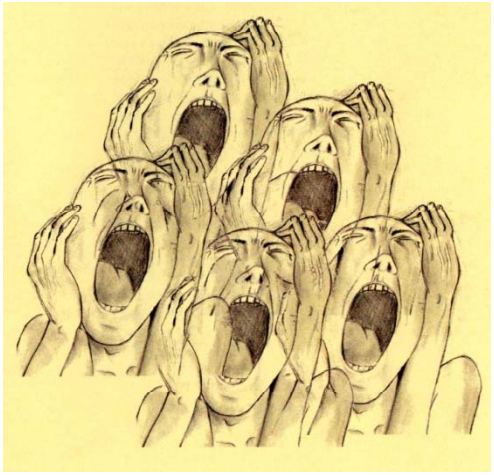
Tale ritardo nell'attacco vocale presente in alcuni tipi di italiano (cfr. /tetto, tutto, petto/ nella città di Catanzaro [tEt:ⁿo ·tut:ⁿo ·pEt:ⁿo]) può essere facilmente descritto con la figura seguente:



Stato delle corde vocali durante la produzione di /tetto/ [tEt:ⁿo] con aspirazione.

La stima puntuale

Supponiamo di essere interessati a determinare la **lunghezza media** del */t/ VOT* in un particolare gruppo: donne adulte residenti in una determinata località



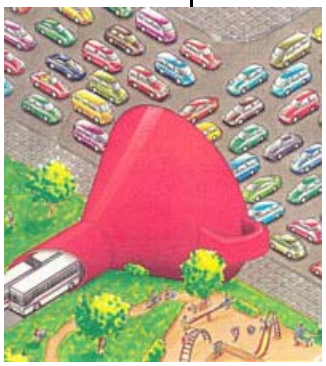
Una misura esatta, a meno di errori grossolani, si può ottenere rilevando il */t/ VOT* per ognuna delle donna adulta.

E' una tale operazione realizzabile in tempi brevi e costi contenuti? Ovviamente no!!!!

L'idea è quella di rilevare il */t/ VOT* su un campione rappresentativo di donne e utilizzare tale informazione campionaria per risalire al */t/ VOT* dell'intera popolazione di donne adulte.

La stima puntuale

L'informazione acquisita tramite il campione viene sintetizzata attraverso una particolare funzione matematica dei dati campionari. Tale funzione è detta STIMATORE

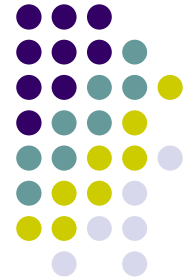


Nel nostro caso lo stimatore da utilizzare sarà la **media campionaria**

Prima di estrarre il campione, la stimatore è una **variabile casuale** in quanto può assumere un qualsiasi valore in un determinato intervallo con una prefissata probabilità. → *distribuzione campionaria*

Selezionato il campione , lo stimatore assumerà un unico valore detto **STIMA**

La stima puntuale

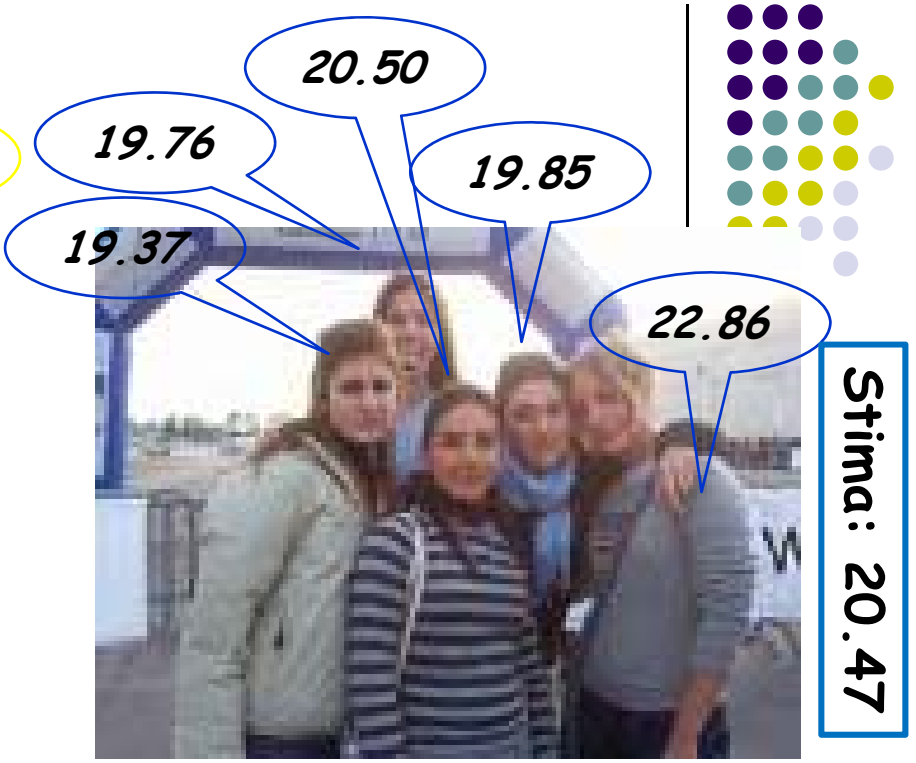


Supponiamo di considerare una popolazione di 50 donne e di rilevare per ciascuna di esse la lunghezza (in ms) del */t/ VOT*

18.27	24.76	19.25	20.46	21.18	23.43	18.40	22.67	13.58	22.12
13.34	24.76	22.90	24.27	14.66	25.02	22.76	24.76	21.03	20.88
20.50	19.85	17.65	20.24	22.86	13.63	23.26	15.19	15.77	16.31
21.15	21.31	28.73	19.62	26.49	14.24	22.85	19.92	25.66	11.32
15.41	20.70	19.45	16.67	17.23	22.29	25.16	19.37	16.78	19.76

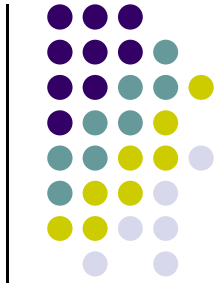
La lunghezza media (μ) del /t/ VOT è pari a **20.19** ms con un deviazione standard di **3.86** ms

Supponiamo ora di non conoscere la popolazione e di volere stimare μ sulla base di uno dei possibili 50^5 = 312 500 000 campioni formati da 5 donne



In nessuno dei tre campioni le stime coincidono con il vero valore della popolazione e solo in uno la stima sembra plausibile.

L'accuratezza delle stime



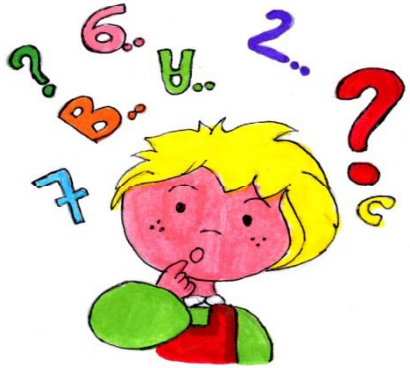
Il vero problema è che non conosciamo il parametro da stimare e, quindi, nessuno giudizio circa l'attendibilità della stima può essere formulato sulla base di un campione osservato.

L'accuratezza della stima può essere valutata solo sulla base delle proprietà statistiche di cui gode lo stimatore:

- **Correttezza:** mediamente le stime coincidono con il parametro incognito da stimare. La media va calcolata su tutti i possibili campioni
- **Consistenza:** all'aumentare della dimensione campionaria la stima si avvicina sempre di più al parametro incognito della popolazione . In altre parole si riduce la variabilità delle stime intorno la parametro

La media campionaria gode di queste proprietà

Per confondere (un po') le idee



Un fenomeno casuale può essere espresso attraverso il semplice modello matematico

$$X = \mu + \varepsilon$$

in cui il valore che esso assume è dato dalla sua media (μ) più un errore un errore casuale (ε). La media è incognita mentre l'errore casuale ha spesso un distribuzione Normale.

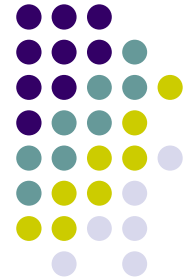
Il problema che si pone è stimare μ sulla base di un campione di n osservazioni, $x_1,\ldots,x_n$

Stimatore media campionaria $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Stima $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Distribuzione campionaria $\bar{X} \approx N\left(0, \frac{\sigma^2}{n}\right)$

Limiti della stima puntuale



- Uno stimatore fornisce un possibile valore del parametro incognito della popolazione. Tale valore cambia al variare del campione selezionato e non si tiene conto della variabilità campionaria
- Stimare esattamente il parametro incognito è impossibile
- In ogni stima comporta un margine di errore che non è possibile misurare . Da qui l'esigenza di presentare accanto alla stima puntuale una qualche misura dell'errore a cui essa è soggetta.

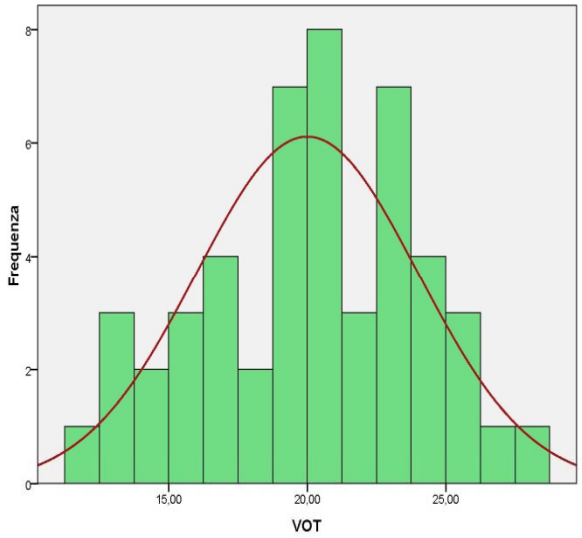
E' preferibile fornire un intervallo di valori intorno alla stima puntuale che offra "sufficienti garanzie" di contenere il valore del parametro.

Il **grado di fiducia** può essere misurato in termini probabilistici.

Intervalli di confidenza per μ

Supponiamo di essere interessati a determinare un **intervallo di valori** all'interno del quale ricada la lunghezza media del \t\ VOT per la popolazione delle 50 donne dell'esempio precedente.

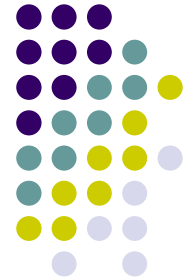
Sull'intera popolazione il VOT presenta un andamento normale con media **20.19** ms e deviazione standard di **3.86** ms



Supponiamo di conoscere la varianza della popolazione ma non la media.

Siamo quindi interessati a determinare un intervallo per la media di una popolazione Normale con varianza incognita

Intervalli di confidenza per μ



La variabile $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ ha una distribuzione Normale

Standard (con media 0 e varianza 1) e

$$P(-1.96 < Z < 1.96) = 0.95$$

Sostituendo a Z la sua espressione ed isolando algebricamente μ otteniamo l'espressione equivalente

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

L'intervallo casuale

$$\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

è detto **intervallo di confidenza (casuale) al 95%** per la media incognita μ

Prima di estrarre il campione, la probabilità che la media incognita sia contenuta nell'intervallo casuale è pari a 0.95

Una volta selezionato il campione, se sostituiamo alla media campionaria la sua stima, otteniamo **l'intervallo di confidenza osservato al 95%**

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\,,\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

$$\bar{x} = 20.47$$

$$\left[20.47 - 1.96\frac{3.86}{\sqrt{5}}\,,\ 20.47 + 1.96\frac{3.86}{\sqrt{5}}\right] = [17.08\,,\ 23.85]$$

L'IC al 95% contiene il valore delle media (μ = 20.19).

Dire che l'intervallo calcolato contiene il vero valore della media con probabilità pari a 0.95 è ERRATO. Essa infatti sarà 0 oppure 1 a seconda se la media è contenuta o meno nell'intervallo.
Il problema è che non conosciamo il vero valore della media e, quindi, non sapremo mai se l'IC contiene o meno tale valore.

Tuttavia, abbiamo un elevato grado di fiducia che l'IC contenga la media incognita. Fatto 100 il massimo grado di fiducia (certezza), quello relativo all'intervallo è pari a 95… ovvero, siamo certi al 95% che l'IC contenga la media

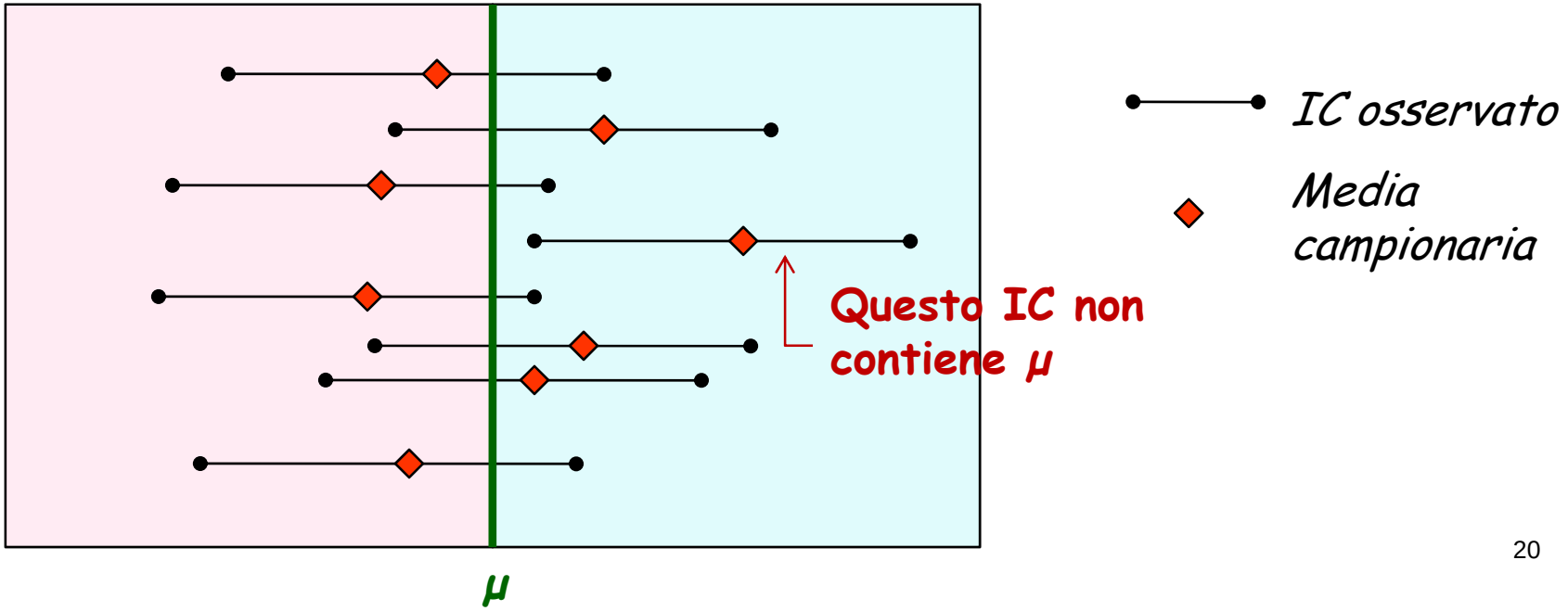
Tale "fiducia" deriva dalla logica sottostante la costruzione degli IC: se estraessimo un numero elevato di campioni, e per ognuno di questi calcolassimo l'intervallo di confidenza, il 95% di questi conterrebbe la media incognita.

… **confidiamo**, quindi, che il nostro intervallo rientri in quel 95% di intervalli che contengono le media

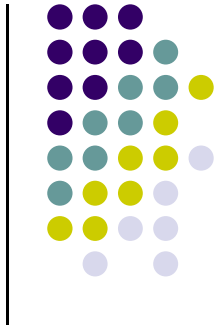
*… **confidiamo**, quindi, che il nostro intervallo rientri in quel 95% di intervalli che contengono le media.*

L'IC calcolato sul campione di veline determina un IC che rientra in quel 5% di intervalli che non contengono la media. Infatti

$$IC = \left[24.06 - 1.96 \frac{3.86}{\sqrt{5}} , 24.06 + 1.96 \frac{3.86}{\sqrt{5}} \right] = [20.67 , 27.44]$$



IC per la media, σ nota



- IC al 90% $\longrightarrow \left[\bar{X} - 1.65\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.65\frac{\sigma}{\sqrt{n}}\right]$
- IC al 99% $\longrightarrow \left[\bar{X} - 2.576\frac{\sigma}{\sqrt{n}}, \bar{X} + 2.576\frac{\sigma}{\sqrt{n}}\right]$
- IC al (1-α)% $\longrightarrow \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Dove $z_{\alpha/2}$ è il percentile al livello (1-α/2)% della Normale Standard

$$z_{\alpha/2} : P(Z > z_{\alpha/2}) = \alpha/2$$

IC e precisione

La precisione di un IC fa riferimento alla sua ampiezza L

$$L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

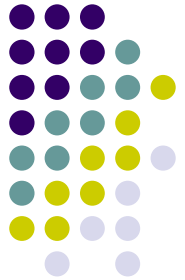
Maggiore è *L*, minore sarà la precisione dell'IC

E' facile fare centro quando il bersaglio è molto grande … ma è troppo banale!

La precisione dipende da tre fattori

- Livello di confidenza (1-α)
- Deviazione standard σ
- Ampiezza campionaria n

IC e precisione



- *Ceteris paribus*, la precisione diminuisce all'aumentare del livello di confidenza
- *Ceteris paribus*, la precisione diminuisce all'aumentare della variabilità
- *Ceteris paribus*, la precisione aumenta all'aumentare della dimensione campionaria

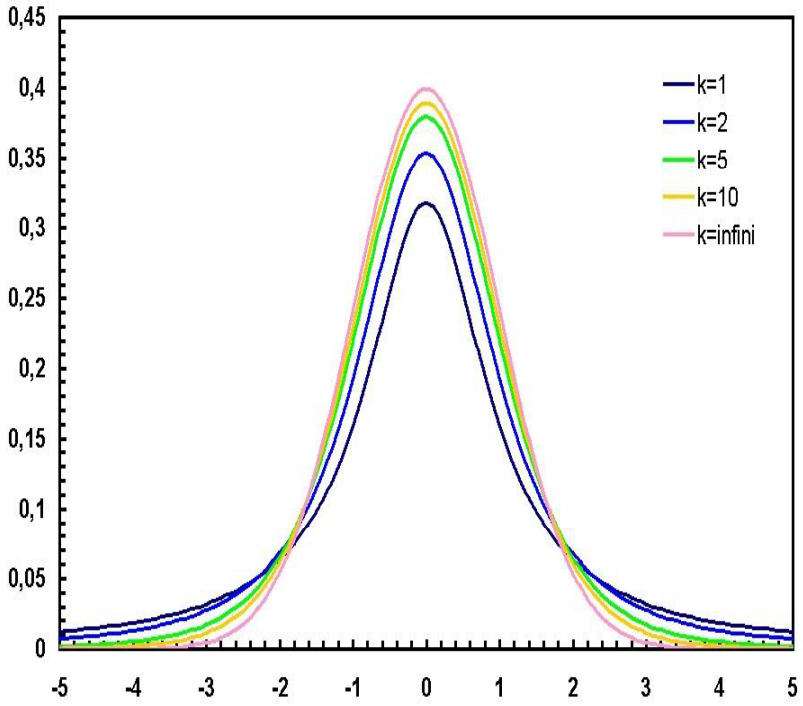
Fissati (1-α) e σ è possibile determinare un'ampiezza campionaria che garantisce un determinato livello di precisione L^*

$$n = \left(\frac{2 z_{\alpha/2} \sigma}{L^*} \right)^2$$

IC per μ, σ incognita

La variabile $T = \dfrac{\bar{X} - \mu}{S / \sqrt{n}}$ ha una distribuzione *t*-Student

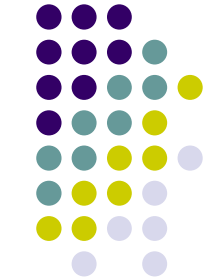
con (n-1) gradi di libertà dove $S = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$



La v.c. *t*-Student presenta alcune caratteristiche:

- ha una forma campanulare e simmetrica centrata sullo zero
- è più "piatta/panciuta" della Normale
- tende alla Normale all'aumentare dei gradi di libertà
- è utilizzata in procedure inferenziali quando la varianza è incognita

IC per la media, σ incognita



Valore di S nel campione

- IC al (1-α)% $\longrightarrow \left[\bar{X} - t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}}\right]$

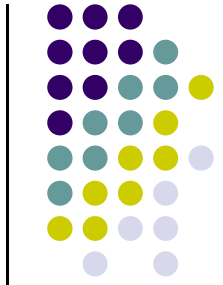
dove $t_{\alpha/2}^{(n-1)}$ è il percentile al livello (1-α/2)% della *t*-Student con (n-1) gradi di libertà

Ad esempio, per il campione di teen-agers, l'IC al 95% per μ è:

$$\left[20.47 - 2.78\frac{1.39}{\sqrt{5}}, 20.47 + 2.78\frac{1.39}{\sqrt{5}}\right] = [18.74, 22.19]$$

L'interpretazione è analoga al caso in cui σ è nota.

IC per una proporzione



Accade spesso di voler stimare la "prevalenza" di un certo attributo nella popolazione oggetto di studio.

Esempio. Un gruppo di ricercatori vuole ottenere una stima della proporzione (p) di bambini in età prescolare affetti da difetti dell'apprendimento.

A tal fine viene condotto uno studio su 200 bambini rilevando che 17 di essi prestano il disturbo.

La stima puntuale è data dalla proporzione campionaria: $\longrightarrow \hat{p} = \dfrac{\text{num. casi favorevoli}}{\text{num. casi possibili}}$

$$\hat{p} = \frac{17}{200} = 0.085 \ (8.5\%)$$

Per campioni di ampiezza sufficientemente elevata, l'IC al livello (1-α)% per p è:

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Così, ad esempio, l'IC al 95% per la proporzione di bambini affetti da disturbi dell'apprendimento è:

$$\left[0.085 - 1.96\sqrt{\frac{0.085(1-0.085)}{200}}\ ,\ 0.085 + 1.96\sqrt{\frac{0.085(1-0.085)}{200}}\right]$$

$$= [0.046\ ,\ 0.124]$$